JADT 2022

PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT 2022
PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA
Volume 2
Michelangelo MISURACA, Germana SCEPI, Maria SPANO 2022

2022 Vadistat press in coedizione Edizioni Erranti

Associazione VADISTAT - Per Simona Balbi
Dipartimento di Scienze Economiche e Statistiche
Università di Napoli "Federico II"
Complesso Universitario Monte Sant'Angelo
Cupa Nuova Cintia 21 - Napoli

Edizioni Erranti
Editoria Indipendente - Libero Sapere
Via Caloprese 23 - Cosenza

www.edizionierranti.org
info@edizionierranti.org

Isbn 979-12-80153-31-9

Edited by
Michelangelo MISURACA
Germana SCEPI
Maria SPANO

JADT 2022

PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Naples, Italy - July 06-08, 2022)

VOLUME 2

VADISTAT PRESS
2022

The editing process of the proceedings has been actively supported by:
Alessandra Belfiore, Nicola d'Alesio, Luca D'Aniello, Agostino Gnasso

Scientific Committee

Ramón Álvarez Esteban (Univ. de León, ES)
Valérie Beaudouin (Telecom ParisTech, FR)
Mónica Bécue (Univ. Polyt. de Catalunya, ES)
Sergio Bolasco (Univ. di Roma La Sapienza, IT)
Lou Burnard (Oxford Univ., UK)
Isabella Chiari (Univ. di Roma La Sapienza, IT)
Anne Dister (Univ. Saint-Louis / UCLouvain, BE)
Jules Duchastel (UQÀM, CA)
Serge Fleury (Univ. Sorbonne Nouvelle - Paris 3, FR)
Cédrick Fairon (UCLouvain, BE)
Serge Heiden (ENS de Lyon, FR)
Domenica Fioredistella Iezzi (Univ. di Roma Tor Vergata, IT)
Margareta Kastberg (Univ de Franche Comté, FR)
Ludovic Lebart (CNRS / ENST, FR)
Jean-Marc Leblanc (Univ. de Créteil, FR)
Alain Lelu (Univ. de Franche Comté, FR)
Dominique Longrée (Univ. de Liège, BE)
Véronique Magri (Univ. Côte d'Azur / CNRS BCL, FR)
Pascal Marchand (Univ. de Toulouse, FR)
Damon Mayaffre (Univ. Côte d'Azur / CNRS BCL, FR)
Sylvie Mellet (CNRS, FR)
Michelangelo Misuraca (Univ. della Calabria, IT)
Denis Monière (Univ. de Montréal, CA)
Bénédicte Pincemin (CNRS, FR)
Céline Poudat (Univ. Côte d'Azur / CNRS BCL, FR)
Gérald Purnelle (Univ. de Liège, BE)
Elias Rizkallah (UQÀM, CA)
Pierre Ratinaud (Univ. de Toulouse, FR)
Max Reinert (CNRS / Univ. Paris-Saclay, FR)
André Salem (Univ. Sorbonne Nouvelle - Paris 3, FR)
Arjuna Tuzzi (Univ. di Padova, IT)
Mathieu Valette (Inalco, FR)
Jean-Marie Viprey (Univ de Franche Comté, FR)

Program Committee

Michelangelo Misuraca (Univ. della Calabria, IT) [CHAIR]
Enrica Amaturo (Univ. di Napoli Federico II, IT)
Laura Antonucci (Univ. di Foggia, IT)
Massimo Aria (Univ. di Napoli Federico II, IT)
Andrea Fronzetti Colladon (Univ. di Perugia, IT)
Fiorenza Deriu (Univ. di Roma La Sapienza, IT)
Giuseppe Giordano (Univ. di Salerno, IT)
Maria Gabriella Grassia (Univ. di Napoli Federico II, IT)
Francesca Greco (Univ. di Roma La Sapienza, IT)
Francesca Della Ratta (INAPP, IT)
Domenica Fioredistella Iezzi (Univ. di Roma Tor Vergata, IT)
Marina Marino (Univ. di Napoli Federico II, IT)
Mario Monteleone (Univ. di Salerno, IT)
Germana Scepi (Univ. di Napoli Federico II, IT)
Maria Spano (Univ. di Napoli Federico II, IT)
Arjuna Tuzzi (Univ. di Padova, IT)

Organising Committee

Maria Spano (Univ. di Napoli Federico II, IT) [CHAIR]
Alessandra Belfiore (Univ. della Campania Luigi Vanvitelli, IT)
Enrico Cafaro (Univ. di Napoli Federico II, IT)
Luigi Celardo (Univ. di Napoli Federico II, IT)
Luca D'Aniello (Univ. di Napoli Federico II, IT)
Nicola d'Alesio (Univ. della Campania Luigi Vanvitelli, IT)
Agostino Gnasso (Univ. di Napoli Federico II, IT)
Giorgio Infante (VADISTAT Per Simona Balbi, IT)
Rocco Mazza (Univ. di Napoli Federico II, IT)
Raffaele Mattera (Univ. di Napoli Federico II, IT)
Agostino Stavolo (Univ. di Napoli Federico II, IT)

TABLE OF CONTENTS

Introduction **pag. xvii**

Invited Speakers

Manuel J. COBO **1**
Bibliometric and science mapping analysis: models, software tools and future challenges

Stefano Maria IACUS **7**
Sentiment Analysis, Social Media and Subjective Well-Being

Max SILBERZTEIN **12**
Linguistic Resources for Corpus Processing: the ATISHS project

Contributions

Juan Abasolo, Naia Eguskiza, Aitor Iglesias **30**
Validación de léxico en lengua vasca mediante análisis lexicométrico de un corpus paralelo multilingüe

Juan Abasolo, Naia Eguskiza, Aitor Iglesias **37**
Construcción de un léxico en lengua vasca para su utilización con el software IRaMuTeQ

Suania Acampa **45**
A hybrid approach to opinion mining in Facebook's disinformative echo chambers

Touria Ait El Mekki, Bénédicte Grailles, Tsanta Randriatsitohaina **52**
Création d'une base de connaissances à partir des messageries spécialisées pour améliorer l'exploration et l'archivage des méls

Ramón Alvarez-Esteban, Mónica Bécue-Bertaut **60**
Order constrained clustering with local stopping rules. Application in textual analysis

Massimo Aria, Corrado Cuccurullo, Luca D'Aniello, Michelangelo Misuraca, Maria Spano **67**
Text Summarization of a scientific document: a comparison of Extractive unsupervised methods

Valerio Basile **74**
Towards Automatic Screening for Fibromyalgia in Italian Social Media Users

Federica Beghini **80**
Stylométrie, ADT et deep learning. Une étude de cas sur la prose romanesque de Milan Kundera

Alessandra Belfiore, Maria Spano, Katerina Mandenaki, Walter Giordano **88**
Exploring Companies' communication strategies in the Covid-19 era

Alessandra Belfiore, Agostino Gnasso, Corrado Cuccurullo, Massimo Aria **95**
AI and ML in accounting and finance: A bibliometric review
Loredana Bellantuono **102**
Complex networks and natural language processing to unfold the social enterprises environment
Pietro Belloni, Margherita Silan, Giulia Cuman **104**
Fake news spreading and sentiment of Italians during the first COVID-19 lockdown
Marion Bendinelli **111**
A Preparatory Corpus-Based Study of Swiss Touristic Brochures. Euphoric markers and Phraseology
Abdallah Benkadja, Ismaïl Biskri, Nadia Ghazzali **119**
Hybride CNN-SVM : vers une meilleure classification des données manuscrites
Valentina Biagioli, Francesca Greco, Giuseppina Spitaletta, Annachiara Liburdi, Rachele Mascolo, Orsola Gawronski, Riccardo Ricci, Emanuela Tiozzo, Ercole Vellone, Teresa Grimaldi Capitello, Michele Salata, Massimiliano Raponi, Immacolata Dall'Oglio **127**
Emotional Text-Mining in Pediatria: esplorazione del self-care nei bambini e nei giovani in condizioni croniche complesse
Magali Bigey **134**
Impact of the expression of feelings and emotions in fake news and their reception: Specific electronic dictionaries and automatic recognition with NooJ
Marco Bolpagni, Marco Broglio, Andrea Innocenzi, Tommaso Ulivieri **141**
EMOtivo: a classifier for emotion detection of Italian texts trained on a self-labelled corpus
Barbara Boschetto, Francesca Della Ratta, Federica Pintaldi, Maria Elena Pontecorvo, Alessia Sabbatini **148**
"Altro specificare". Effetto del Covid-19 sulla codifica dei campi "altro" nell'indagine Istat sulle forze di lavoro
Rosalie Bourdages, Roxane Gagnon, Denis Foucambert **156**
Évaluation et description de récits oraux spontanés d'élèves du primaire : pour l'utilisation des méthodes mixtes en sciences de l'éducation
Camille Bouzereau **164**
Complémentarité du quantitatif et du qualitatif pour l'analyse de données textuelles. Analyse du discours du Front National (2000-2017)
Giulia Bracco, Maria De Martino, Alessandro Laudanna **172**
Semantic and morphological properties of gender assignment in Italian: distributional data on common nouns

Andrea Briglia, Massimo Mucciardi, Giovanni Pirrotta **180**
The development of word frequency distribution in first language acquisition. An analysis on a spoken language corpus of French children
Joanna Byszuk, Quinn Dombrowski **188**
Stylometric investigations into translationese: The Baby-Sitters Club across languages
Stefania Capogna, Giulia Cecchini, Maria Chiara De Angelis, Francesca Greco, Emanuela Proietti **197**
Analisi comparativa dei rapporti nazionali: il caso del Progetto Erasmus+ ECOLHE
Rosanna Cataldo, Gabriella Punziano, Carmine Barricelli, Gabriele Luciano, Barbara Saracino, Ferdinando Iazzetta **204**
Three main dimensions in the construction of the expert during the Italian Covid-19 vaccination campaign: positional, reputational, and communicational spheres in comparison
Elena Catanese, Monica Scannapieco, Mauro Bruno, Luca Valentino **213**
The Italian Social Mood on Economy Index: recent methodological developments
Radek Čech, Miroslav Kubát, Ján Mačutek, Michaela Koščová **221**
Does an author leave a syntactic footprint ?
Roy Cerqueti, Valerio Ficcadenti, Gurjeet Dhesi, Marcel Ausloos **229**
A methodological strategy for assessing the Markovian stochastic structure of the text analysis-based rank-size laws
Marie Chandelier, Sascha Diwersy **234**
Analyse du profil sémantique des expressions biodiversité et diversité biologique dans la presse écrite française
Giacomo Chiara, Diego Romaioli, Alberta Contarello **243**
Rappresentazioni sociali della migrazione e dell'identità culturale in testi letterari di scrittrici e scrittori di origine africana
Barbara Cordella, Alessandro Gennaro **252**
Alla ricerca del benessere: le famiglie adottive
Rosario D'Agata, Domenico De Stefano, Francesco Santelli **259**
The 'words' of no green pass communities on twitter. A two-mode semantic network analysis
Angela Maria D'Uggento, Albino Biafora, Claudia Marin, Fabio Manca, Massimo Bilancia **266**
Sentiment analysis of emotions in social media during the COVID-19 pandemic
Giovanni De Gasperis, Pasquale Pavone, Sergio Bolasco **274**
La risorsa di Italiano Standard ad alta variabilità linguistica per misurare la peculiarità di un corpus
Alberto Maria De Mascellis, Maria Gabriella Grassia, Agostino Stavolo **282**
Topic modeling and sentiment analysis: an analysis of tweets about the game CyberPunk 2077

Francesca Della Ratta, Monya Ferritti **290**
Gli insegnanti dopo l'anno della didattica a distanza
Alessandro Delmonte, Matteo Farné **298**
Evaluating customer satisfaction through Amazon review analysis : the bluetooth earphones example
Vanina Deneux Le Barh **306**
Du discours sur le travail avec les chevaux aux valeurs communes des professions anthropoéquines
Francesco Paolo Di Candia, Nicoletta Roberto, Domenica Fiordistella Iezzi **314**
Italian Institutional communication in pandemic period: a chronological analysis of Prime Minister speeches
Sami Diaf, Florian Schütze **321**
Estimating Uncertainty in Narrative Economics Using Latent Semantic Scaling
Sami Diaf, Rachid Toumache **328**
Cross-lingual Macroeconomic Narratives: Different Views Of The Same Problem ?
Anne Dister, Hubert Naets, Patrick Watrin **334**
Présence des noms d'agents au féminin dans un corpus de presse. Étude comparée Québec-Belgique
Sascha Diwersy, Ivana Didirková, Christelle Dodane, Fabrice Hirsch, Giancarlo Luxardo **340**
Les temps de la crise sanitaire au prisme d'une série chronologique : une étude phonético-textométrique
Alexandre Dniestrowski, Louis Allard **348**
The "Augmented" Historian
Chuanming Dong, Philippe Gambette, Catherine Dominguès **354**
Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle
Maximiliano Duran **361**
Automatic retrieving of derived Quechua verbs
Maciej Eder **369**
Improving the performance of word frequencies in authorship attribution
Louis Escouflaire, Antonin Descampe, Cédrick Fairon **377**
L'évolution de la subjectivité linguistique dans le journalisme web du XXIe siècle : analyse d'un corpus belge francophone d'articles de 2010 à 2021
Shira Fano, Gianluca Toschi **385**
The economy and the Web: RETI a regional economic twitter index
Manuel Favaro, Marco Biffi, Simonetta Montemagni **392**
Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione

Marco Ferracci, Germana Scepi, Maria Spano, Michelangelo Misuraca **400**
Nodi discorsivi e processi di progettazione condivisa: Milano e Torino
Rosa Ferri, Valentina Bua, Marzia Curia, Alessia D'Avack, Antonino Firetto, Federica Mauro, Francesca Greco **408**
Text Mining e Salute Organizzativa: il caso di Roma Capitale
Marzia Freo, Alessandra Luati **414**
Lasso-based variable selection methods in text regression: the case of short texts
Andrea Fronzetti Colladon **422**
Being known, being known for something and being good: A new textual measure of organizational reputation
Augusto Gamuzza, Francesca Greco, Anna Maria Leonora **424**
Text mining e analisi ermeneutica: il caso delle pratiche di volontariato internazionale
Franco M. T. Gatti **432**
Rehearsals of ideological placement on a corpus of parliamentary questions
Giuseppe Giordano, Maria Carmela Catone **441**
Islands and structural holes for mapping topics in scientific field. An application to Social Research Methodology
Maria Gabriella Grassia, Marina Marino, Rocco Mazza, Agostino Stavolo **446**
Analysis of the public debate on DDL Zan on Twitter: an application of the Structural Topic Model
Deborah Grbac **453**
From meta-data to research-data. How to use text analysis techniques to do Information retrieval as a propaedeutic practice to Data literacy
Francesca Greco, Raffaella Gallo, Alessandro Polli, Fiorenza Deriu **461**
A comparison between machine learning and non-supervised techniques using the GP Index
Francesca Greco, Gevisa La Rocca, Giovanni Boccia Artieri **468**
Sentiment analysis dei corpora multilingua: La percezione pubblica del Covid-19 in Europa
Francesca Greco, Alessandro Polli, Federico Siciliano **475**
Leveraging Deep Learning models to assess the temporal validity of Emotional Text Mining method
Massimo Guarino, Violetta Simonacci, Michele Gallo **482**
Linguistic complexity and engagement in social media communication for cultural heritage

Domenica Fioredistella Iezzi, Roberto Monte **489**
Sales forecast and electronic word of mouth: the power of feelings
Julie-Pier Jean, Sylvia Kasparian **495**
Cree Legends and Narratives au prisme d'IRaMuTeQ
Kristina Kocijan, Krešimir Šojat **503**
Who is Guilty and Who is Responsible in the Croatian Parliament: A Linguistic Approach
Tomasz Kruszewski, Joanna Michalak **511**
Dynamic and content of COVID-19 related tweets during early stage of pandemic
Daniela Laricchiuta, Andrea Termine, Carlo Fabrizio, Noemi Passarello, Francesca Greco, Fabrizio Piras, Eleonora Picerni, Debora Cutuli, Andrea Marini, Laura Mandolesi, Gianfranco Spalletta, Laura Petrosini **519**
The affecting meaning of words: multidimensional analysis of emotional communication
Sara Laurita, Rossella Nicoletti, Alfredo Fortunato, Carmelofrancesco Origlia **527**
Framing the presence of tourism policies in the Calabrian public agenda during Covid-19 crisis: analysis of online news from local newspapers using 3-wave datasets
Ludovic Lebart **536**
Poems and Songs. What can Textual Data Analysis do?
Virginie Lethier, Emilie Née, Hugo Dumoulin **544**
Caractériser les logiques disciplinaires et institutionnelles dans les rapports d'autoévaluation
Daniele Licari, Maria Francesca Romano, Giovanni Comandé **552**
Automatic Anonymization of Italian Legal Textual Documents using Deep Learning
Rosaria Lombardo, Matteo Pascoli, Nicola Grandi **560**
Text Mining and Variants of Correspondence Analysis for analysing written Italian of University students
Carlos Maciel, Damon Mayaffre, Laurent Vanni **568**
Corpus non alignés et ADT. Essai de comparaison entre les présidents français et brésiliens de l'ère contemporaine
Marina Marino, Rocco Mazza, Gabriele Luciano **576**
Online communication and political participation: a study dedicated to the referendum Cannabis Legale
Raffaele Mattera, Michelangelo Misuraca, Germana Scepi, Maria Spano **584**
Clustering of financial time series: a bibliometric analysis

Damon Mayaffre **591**
Contribution à l'histoire d'un concept : occcurrences et cooccurrences de 'souveraineté' dans les discours de Charles de Gaulle (1958-1969) et d'Emmanuel Macron (2017-2022)
Simona Mercurio **599**
What about corruption? A text analytics method for literature review
Silvia Monaco, Anna Cortellino, Francesca Greco, Michela Di Trani **607**
Using Emotional Text Mining to explore the Spanish culture of Organ Donation
Cyrielle Montrichard **613**
Questionner la posture énonciative d'un journal à visée humoristique: étude textométrique et contrastive
Jean Moscarola, Florence Laval, Caroline Mothe **621**
Recherche Qualitative : Analyse lexicale et sémantique de grands corpus d'entretiens ou de documents
Alejandro J. Napolitano Jawerbaum, Patrick Juola **629**
Co-authorship in Carmen Mola, a case study
Luisa Orrù, Marco Cuccarini, Christian Moro, Monia Paita, Davide Bassi, Giovanni Da San Martino, Nicolò Navarin, Gian Piero Turchi **636**
From fake news identification to news persuasion level study: the application of a machine learning model with the M.A.D.I.T. methodology
Luisa Orrù, Jessica Neri, Marco Cuccarini, Antonio Iudici, Nicolò Navarin, Gian Piero Turchi **644**
Biographical change and Machine Learning with MADIT: a predictive model proposal for data and intervention analysis
Aylin Pamuksaç **651**
Similarities and disparities among French modal verbs through co-occurrence correspondence analysis
Daniel Pélissier, Jérôme Bousquié **659**
Le temps dans le discours, expérimentation d'un protocole d'observation des caractéristiques temporelles d'un corpus d'avis de salariés
Manon Pengam, Agata Jackiewicz, Pascal Marchand **667**
Saisir par la textométrie la notion de radicalisation dans des corpus institutionnels hétérogènes
Marie Pérès, Jean-Marc Leblanc **675**
Réalité virtuelle: quels apports pour la visualisation en textométrie/lexicométrie ?
Loredana Piervisani, Mariachiara Figura, Paola Arcadi, Angela Durante, Ercole Vellone, Rosaria Alvaro **683**
Vocazione e professione: l'imperativo morale dell'infermiera italiana nel periodo fascista

Bénédicte Pincemin, Serge Heiden, Franck Mazuet **691**
The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection
Sophie Piron **699**
Stylométrie des Elemens de la grammaire Françoise (1780) ou comment Lhomond a supplanté ses prédécesseurs
Lucia Ráčková **708**
Les erreurs interférentielles et le portfolio linguistique du locuteur, deux variables corrélées par le rho de Spearman
David Reymond, Manuel Durand-Barthez, Heman Khouilla, Sandrine Wolff **712**
Retrouver l'inventeur-auteur: la levée d'homonymies d'autorat entre les brevets et les publications scientifiques
Annabel Richeton, Virginie Lethier, Margareta Kastberg Sjöblom **720**
La stabilité du cadre discursif du Ministère des Affaires Etrangères à l'épreuve de la textométrie
Valentina Rizzoli, Andrea Azzoni, Giulia Rivellini, Arjuna Tuzzi **728**
Classificazione manuale vs automatica dei resoconti delle chiamate ricevute da Telefono Amico Italia (2016 -2020)
Valentina Rizzoli, Laura Soledad Norton, Alessandro Meneghini, Mauro Sarrica **736**
How far is the risk? Detecting socio-psychological processes from consciousness to denial in social media exchanges
Valentina Rizzoli, Matilde Trevisani, Arjuna Tuzzi **744**
Life cycle of ideas in the Journal of Personality and Social Psychology: A history of US social psychology
Minerva Rojas **751**
Analyse contrastive interlangue des marqueurs discursifs en Français Langue Étrangère et en Français Langue Maternelle
Maria Francesca Romano, Giovanni De Gasperis, Pasquale Pavone, Sergio Bolasco **758**
Potenzialità di TaLTaC nella anonimizzazione di sentenze della magistratura
Roberto Rondinelli, Stefano Marmani, Valerio Ficcadenti **765**
An analysis of the Jesus' community in the Gospel of Matthew, Mark, Luke, John, and Acts of Apostles networks
Corinne Rossari, Cyrielle Montrichard, Claudia Ricci, Linda Sanvido **772**
Subjectivity and connectors: study of the parameters of discourse genres and languages
Simona Ruggia, Minerva Rojas **780**
Les marqueurs discursifs : étude comparative de leur utilisation et fréquence chez les locuteurs francophones et dans les manuels de français langue étrangère

Camilla Salvatore, Silvia Biffignandi, Annamaria Bianchi **787**
Assessing Corporate Social Responsibility Communication on Twitter: A Composite Indicator Approach
Sara Santilli, Stefano Sbalchiero **794**
The use of IRaMuTeQ for longitudinal analysis. A study on postgraduate course participants' reflections..
Andrea Sciandra **801**
Dimensionality Reduction of Unstructured and Network Data for Stance Detection
Mariangela Sciandra, Alessandro Albano **809**
A two-stage LDA algorithm for ranking induced topic readability
Ludovica Segneri, Andrea Fronzetti Colladon, Claudia Fabiani, Anna Laura Pisello **815**
Text mining for socially responsible behavior: fostering the adoption of green technologies
Gerardo Sierra, Juan-Manuel Torres-Moreno, Mercè Lorente **817**
Exploration de contextes définitoires en français : cas d'étude sur la COVID-19
Andrea Simonetti, Nicoletta D'Angelo, Giada Adelfio **827**
Marked Hawkes processes for Twitter data
Valérie Thon, Laurent Vanni, Dominique Longrée **834**
Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison: à propos de l'étude stylistique des lettres de Pierre Damien
Alice Tontodimamma, Stefano Anzani, Valerio Basile, Marco Stranisci, Lara Fontanella **842**
Comparison of two annotation schemes to derive offensiveness scores in HurtLex
Gian Piero Turchi, Luisa Orrù, Annalisa Trovò, Christian Moro, Antonio Iudici, Nicolò Navarin **850**
Introducing Dialogic Process Analysis in Natural Language Processing
Laurent Vanni, Magali Guaresi, Véronique Magri **858**
Convolution et marqueurs multidimensionnels. Description des représentations genrées dans un corpus de films français
Pierre Wavresky **866**
Crise du Covid et ressenti des acteurs du secteur social et médico-social : une analyse textométrique de journaux de bord et d'entretiens
Emma Zavarrone, Alessia Forciniti, Martha Friel **874**
Exploratory textual network for Italian Third Mission
Emma Zavarrone, Alessia Forciniti, Marta Muscariello **879**
SustDict: a customized lexical-based dictionary for CSR
Lichao Zhu, Fabrice Issac, Touria Ait El Mekki, Olivier Hû **887**
Construction of an encyclopedic dictionary of tourism from scientific corpus

INTRODUCTION

The *International Conference on the Statistical Analysis of Textual Data* (JADT, Journées d'Analyse Statistique des Données Textuelles) has been at its 16th occurrence. It was held for the first time in Naples, from the 6th to the 8th of July 2022, organised by the VADISTAT Per Simona Balbi Association at the University of Naples Federico II. This biennial conference has continuously gained importance since the first edition in Barcelone (1992), and with the editions of Montpellier (1994), Rome (1996), Nice (1998), Lausanne (2000), Saint-Malo (2002), Louvain-la Neuve (2004), Besançon (2006), Lyon (2008), Rome (2010), Liège (2012), Paris (2014), Nice (2016), Rome (2018), Toulose (2020). The JADT conference presents, every two years, the state of the art concerning theories, problems, methods and applications in several domains, sharing a quantitative approach to the study of lexical, textual, pragmatic or discursive features of information expressed in natural language.

In this edition, we decided to continue the innovation process of the traditional conference structure. After the successful experience of JADT2018, the papers were limited to a maximum of 6 pages. The idea is that the contributions were not yet in their final version, and the exchange with other scholars during the conference led to their improvement. A selection of extended papers presented at the JADT2022 conference will be published, after another reviewing process, in a book published by Springer and in several special issues of acknowledged Journals. The perspective of enhancing the papers discussed during the conference will allow the JADT community to keep the network of contacts and lively exchanges active. Moreover, we tried to involve young researchers by offering two awards sponsored by VADISTAT - Per Simona Balbi, following the attitude of our beloved Simona in supporting students and researchers in training interested in the Statistical Analysis of Textual Data. All our efforts to organise a great scientific meeting are dedicated to her memory.

These proceedings collect 118 contributions delivered by more than 300 scholars from 15 countries spread all over the world. The papers include issues open to all scholars and researchers working in the domain of textual data analysis, ranging from lexicography to text mining, information science, analysis of political discourses, content analysis and computational linguistics. The invited speakers focused on central issues for textual data analysis, presenting open and new research frontiers concerning linguistic resources for corpus processing, social media mining and science mapping. The proceedings follow an alphabetical order by the surname of the first author of the contributions.

Michelangelo Misuraca, Germana Scepi, Maria Spano

ACKNOWLEDGEMENTS

We want to express our thanks to the 50 reviewers that offered their assistance in selecting and reviewing the papers published in this volume: Ramón Álvarez Esteban, Enrica Amaturo, Massimo Aria, Mónica Bécue-Bertaut, Rosanna Cataldo, Isabella Chiari, Francesca Della Ratta, Fiorenza Deriu, Anne Dister, Jules Duchastel, Cédrick Fairon, Annick Farina, Andrea Fronzetti-Colladon, Giuseppe Giordano, Maria Gabriella Grassia, Francesca Greco, Serge Heiden, D. Fioredistella Iezzi, Sylvia Kasparian, Alexei Lavrentev, Ludovic Lebart, Jean-Marc Leblanc, Alain Lelu, Dominique Longrée, Véronique Magri, Pascal Marchand, Marina Marino, Raffaele Mattera, Damon Mayaffre, Rocco Mazza, Michelangelo Misuraca, Mario Monteleone, Pasquale Pavone, Bénédicte Pincemin, Alessandro Polli, Céline Poudat, Ilaria Primerano, Pierre Ratinaud, Maria Francesca Romano, Germana Scepi, Andrea Sciandra, Ludovica Segneri, Federico Siciliano, Max Silberztein, Maria Spano, Matilde Trevisani, Arjuna Tuzzi, Laurent Vanni, Emma Zavarrone.

JADT2022 was organised under the patronage of SIS (Sociatà Italiana di Statistica), SIEDS (Società Italiana di Economia Demografia e Statistica), AIS-Metodologia (Associazione Italiana di Sociologia - Sez. Metodologia), ILC-CNR (Istituto di Linguistica Computazionale - Consiglio Nazionale delle Ricerche), ISTAT (Istituto Nazionale di Statistica). We are very grateful to the sponsor, PRISMA s.r.l., and all the contributors that allow us to carry out the conference: DISES - Università Federico II di Napoli, DSS - Università Federico II di Napoli, DISPS - Università di Salerno, ASA (Associazione per la Statistica Applicata), Fondazione Saccone, MDPI - Machine Learning & Knowledge Extraction, K-Synth.

As regards the organisation of the conference, we want to thank all the members of the local committee: Maria Spano, Alessandra Belfiore, Enrico Cafaro, Luigi Celardo, Luca D'Aniello, Nicola d'Alesio, Agostino Gnasso, Giorgio Infante, Rocco Mazza, Raffaele Mattera, Agostino Stavolo.

Special appreciations also go to Massimo Aria, Andrea Fronzetti-Colladon, Francesca Della Ratta & Fiorenza Deriu, Giuseppe Giordano & Ilaria Primerano, Francesca Greco, Marina Marino, Mario Monteleone, for the organisation of special tracks concerning Science Mapping, Text Mining for Business, Integrated methods for information extraction, Network tools, Text Mining for Health, Dimensionality reduction and NLP rule-based environments.

Islands and structural holes for mapping topics in scientific field. An application to Social Research Methodology

Giuseppe Giordano[1], Maria Carmela Catone[2]

[1]University of Salerno, Italy – ggiordano@unisa.it
[2]University of Barcelona, Spain – mcatone@ub.edu

Abstract
Scientometrics analysis allows to map the dominant characteristics of a scientific field with different aims and techniques. In this paper we propose a method to characterize topics exploiting some network topological features derived from text co-occurrences in bibliometric data. We mainly refer to keywords co-occurrence networks, even different sources of texts in a scientific document, such as title or abstract, can be used. Textual units, as nodes of the networks, may show different patterns, as cohesive subgraphs, isolates, as well bridging network components.
Interpreting roles and positions of nodes in the framework of keywords co-occurrence networks, gives a substantive theoretical background to a modified version of the Cobo's Strategic Diagram.
The proposed method is aimed at highlighting the role of brokerage nodes, coherently with the theoretical background of structural holes in innovation. Subgraphs of co-occurrence texts emerging as Islands are interpreted as thematic topics of the scientific fields and characterized by clustering coefficient and Closeness centrality indices used as indicators of the role and positions of the emerging thematic topics with respect to the whole network.

Keywords: Cluster Coefficient, Cohesive Subgraph, Co-occurrence keywords, Network Analysis, Strategic Diagram

1. Introduction

Scientometrics studies allow to map the dominant characteristics of a scientific field with different aims and methods. The two main goals are to evaluate the performance of scientific communities and exploring trends and facts of a specific domain at micro (individual scholar), meso (research center, academic department, university), and macro level (the whole domain). In this work, we refer to the last thread of studies, in order to contribute at the characterization of a scientific field, detecting specific themes as potential engines of innovations,

as they act as structural holes (Burt, 2004).
In recent years, we observed a growth of specialized studies in a wide range of scientific domains, affirming the multidisciplinary of Scientometrics studies. Social science, engineering, medical, health sciences, and environmental sciences are from the leading fields that most contribute to the growth of the Scientometrics field.
The common informative sources are bibliometric data and several analytical tools have been realized to extract knowledge from scientific publications. The type and the quality of information depending on the different aims, but the availability of large datasets, through the Internet, led to the development of new techniques and to the growth of the network approach during the first decade of this century.
In this paper we exploit the advantage to extract bibliometric data, in a relatively easy way, to retrieve network data in terms of textual co-occurrences, inducing common concepts in documents (Aria and Cuccurullo, 2017). Textual data can be obtained in forms of authors' keywords, document title or abstract, we limit our study to authors' keywords co-occurrence networks.
As a result of the co-word analysis, a set of detected topics is obtained. In order to represent the results in a visual way, different visualization techniques can be used. In our approach, we refer to the proposal by Cobo et al. (2011a) and Cobo et al. (2011b) to display results by means of the so called *Strategic Diagram*. The Strategic Diagram is a chart where emerging themes are located according to two network statistics: density and degree centrality.
We aim at giving an alternative definition of the original tool, introducing different network statistics consistent with our research hypotheses. The main idea relies on the role that specific keywords may have because of their position in the co-occurrence network. In analogy to the Structural Hole theory (Burt, 2004), we look for nodes (words) that are in the middle of cohesive sub-graphs, not directly connected between them. Such words may act as brokers from two concepts, giving raise to potential engines for innovative themes.
On the other hand, a distinction will be made if the thematic concept appears as a part of the sub-graph characterized by a structural hole, lying in its cohesive part or in its bridging part. The different shadows of brokerage and embedding giving peculiar substantive meaning in terms of life's cycle of the scientific field topics.
This paper is organized as follows. Firstly, motivations and aim are introduced to make coherent hypotheses. Consistently with our hypothesis, we propose an alternative formulation of the strategic diagram, based on specific network statistics, that is the measures of structural holes, clustering coefficients and closeness centrality with respect to the global network. Analogously to the strategic diagram, the interpretative keys will be given. The proposed approach will be applied to analyze the potential innovative areas of the *Social Research Methodology* field.

2. Motivation and aims

The use of words co-occurrence network to describe scientific fields refers to the works by Callon et al. (see Callon et al., 1983). Content analysis technique measures the association between elementary informative items in textual data. Specifically, it relates the common presence of words in documents to define clusters of terms to be interpreted as semantic groups of different topics emerging in the scientific field (Cobo et al, 2011b). The analysis may be conducted in a longitudinal framework, separating the whole time range in consecutive sup-periods and looking for the evolution patterns of the scientific domain. As reported in Cobo et al. (2011b), several visualization techniques have been proposed in the literature in order to provide a way of exploring and suggesting the interpretation of results (Small, 2006). We refer to the Strategic Diagram as described in Callon et al. (1991).

A theme is settled as a subgraph formed by the keywords and their interconnections, it is called a thematic network. To each thematic network is given a substantive meaning and it is labeled using the name of the most significant keyword in the associated theme (usually identified by the most central keyword of the theme). For instance, a typical strategy for building a strategic diagram is based on the following steps: i) collecting raw data from bibliographic sources (that is either ISI Web of Science or Scopus, among others); ii) selecting the kind of information to analyze (Title, Abstract, Author's keywords, Indexed keywords, or combinations of them); iii) extracting the co-occurrence frequency for any pairs of keywords; iv) computing a similarity/dissimilarity measure based on frequency co-occurrences; v) carry out a clustering algorithm to detect subsets of keywords that show internal cohesion and are most separate each other. The detected clusters, having internal high homogeneity in terms of co-occurrence similarity among keywords, represent thematic concepts of interest for the field. These clusters are the thematic (sub-) networks. As final results, a cluster's labeling post-process is required or an automatic labeling algorithm can be used.

However, dealing with keywords networks, or more in general, with textual networks, whereas the interest for cohesive subgroups of words is straightforward, since they should form some meaningful constructs, also the presence of weakly connected words can be of interest. Specifically, the presence of structural holes in a co-word network has to be regarded as a potential signal of emerging latent constructs. To better clarify this proposition, we recall the concept of node-brokerage role and the importance of structural holes in defining innovative constructs (Labun A. and Wittek R., 2014).

Spanning structural holes gives access to new ideas and resources and it is a way to broker a connection between two disconnected parties. Therefore, it should be in the exploration of the holes around and between these thematic networks

where it is also likely to find concepts of interest, bridging the hole to access to innovative topics (Burt, 2004; Cowan and Jonard, 2007).
Some networks rich in structural holes should represents a scientific field where novel valuable ideas could arise by increasing the connections to merge the distinct themes in new ways. Moreover, in pure social networks brokerage appears to provide the opportunity for social "gatekeeping" – *regulating the access of others to the tightly knit group one belongs to, while at the same time controlling the ways in which one's own group members learn about information coming from other groups* (Burt 2004). In analogy, keywords that act as gatekeepers, should play a strategic role as potential bridge to fuse complex concepts or as cut-node, able to separate two distinct themes.
Our hypothesis is that, as in the case of knowledge flows, the existence of keywords that act as *broker* between cohesive subgroups - or *mature theme* in the language of strategic diagram - are the best suited keywords able to induce the emergence of innovative themes. Innovative themes emerge when new ties appear between keywords belonging to different mature themes.
In order to discuss the proposed strategy, we will analyze a real dataset extracted from the Scopus repository. The bibliometric data refer to the field of Social Research Methodology in the time range from 2003 to 2018. The whole dataset refers to thirty years of Social Research Methodology and has been originally discussed in Catone et al. (2020), here the time horizon will be reduced to more recent years for a reliable analysis.

References

Aria, M., Cuccurullo, (2017). bibliometrix: An R-tool for comprehensive science mapping analysis, The R-package 'bibliometrics'.
Burt, R., *Structural holes: The social structure of competition*, Harvard Press, Boston, MA. (1992).
Burt, R., Structural holes and good ideas, *American Journal of Sociology*, 110 (2), 349–399 (2004).
Callon, M., Courtial, J.-P., Turner, W.A., Bauin, S., From translations to problematic networks: An introduction to co-word analysis, *Social Science Information*, 22 (2), 191-235 (1983).
Catone, M. C., Diana, P., Giordano, G. (2020). Keywords Co-occurrence Analysis to Map New Topics and Recent Trends in Social Research Methods. In Barolli L., Amato F., Moscato F., Enokido T., Takizawa M. (eds) *Advanced Information Networking and Applications*. Cham Springer.1078-1088.
Cobo, M.J., Lopez-Herrera, A.G, Herrera-Viedma, E., Herrera, F., Science Mapping Software Tools: Review, analysis, and cooperative study among tools, *Journal of the American Society for Information Science and Technology*, (2011a).
Cobo, M.J., Lopez-Herrera, A.G, Herrera-Viedma, E., Herrera, F., An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical

application to the fuzzy sets theory field, *Journal of Informetrics*, 5 (1), 146-166 (2011b).
Cowan, R., Jonard, N., Structural holes, innovation and the distribution of ideas, *Journal of Economic Interaction and Coordination*, 2 (2), 93–110 (2007).
Guler, A.T., Waaijer, C.J., Mohammed, Y., Palmblad, M., Automating bibliometric analyses using Taverna scientific workflows: A tutorial on integrating Web Services, *Journal of Informetrics*, 10 (3), 830-841 (2016a).
Guler, A.T., Waaijer, C.J., Palmblad, M., Scientific workflows for bibliometrics, *Scientometrics*, 107 (2), 385-398 (2016b).
Labun A., Wittek R., (2014) Structural Holes. In: Alhajj R., Rokne J. (eds) *Encyclopedia of Social Network Analysis and Mining*. Springer, New York, NY.
Salton,G., Buckley, C., Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, 24 (5), 513-523, (1988).
Small, H., Tracking and predicting growth areas in science, *Scientometrics*, 68 (3), 595-610 (2006).

Analysis of the public debate on DDL Zan on Twitter: an application of the Structural Topic Model

Maria Gabriella Grassia, Marina Marino,
Rocco Mazza, Agostino Stavolo

University of Naples Federico II – mariagabriella.grassia@unina.it
marina.marino@unina.it rocco.mazza@unina.it
agostino.stavolo@unina.it

Abstract
Social media play a key role in analysing the public debate on political issues. The amount of information available and the ability to share opinions on various topics has led to an increase in the production of user-generated content. Indeed, data from social media platforms such as Twitter offer new possibilities because it provides a public arena to understand feedback and opinions on certain political topics. In this article, we propose an analysis of the debate on Twitter regarding the DDL Zan, the bill that aims to toughen the penalties of crimes and discrimination against homosexuals, transsexuals, women, and disabled people. We define a strategy to extract the main topics of the public debate considering the role assumed by the so-called influencers within the Twitter community. For this reason, we chose to use the Structural Topic Model (STM) including in the model the covariates describing the user's profiles of interest. This allows to identify the most relevant topics and to assess the impact of covariates in the model to study how covariates impact on them.

Keywords: Structural Topic Model; DDL Zan; Topic Modeling

1. Introduction

Social media plays a key role in analyzing public debate on policy issues, because user-generated contents (Chatziadam et al., 2020) offers a great opportunity to automatically detect opinions and trends on topics of interest. In fact, microblogging platforms such as Twitter are a great way for users to share ideas, opinions, and comments about products, services. Detecting topics in online content facilitates the identification of emerging social trends and the analysis of public reactions to policies (Sobkowicz et al., 2012). Therefore, we propose a strategy to extract the main topics of public debate by considering the role assumed by so-called influencers within the Twitter community. We present a

case study based on the use of the Structural Topic Model (STM) on a debated political event: the "DDL Zan". The bill includes measures to prevent and combat discrimination and violence on the grounds of sex, gender, sexual orientation, gender identity, and disability. The possibility of analysing public discussion on Twitter through a particular type of topic modeling allows to assess the impact of covariates in the model to study how covariates may influence the words representing a topic.

2. Literature Review

Twitter functions as a public sphere because everyone can see tweets, hashtags allow for organization around topics, and retweets allow for rapid dissemination (Marichal et al.,2020). On the platform, the dissemination of information is controlled through influencers (Colleoni et al., 2014). The possibility of sharing opinions and comments through tweets and the ability to interact on the platform becomes an important area of analysis for researchers. Castellò et. al (2021) performed a quantitative analysis of message interactivity as a measure of interactivity: endorsement and amplification. In this context, topic modeling facilitates understanding the conversations between people in online communities, as well as extracting useful patterns and understanding from their interactions in addition to what they share on social media websites (Hamed et al., 2019). It is used for analyzing textual content that uncovers the latent thematic structure in document collections for identifying emerging topics. In the literature, the focus of using topic modeling in recent years has been on the production of short texts on social media. Modeling in this domain is often more challenging due to character limits (e.g., tweets consisting of 280 characters). Structural topic modeling (STM) is a variant of LDA that is designed to precisely address this limitation. For this reason, STM was applied in numerous jobs to analyse online discourse (Kwon et al. 2019; He et al.,2020).

3. Structural Topic Model

The structural topic model (STM) introduced by Roberts et al. (2014) allows us to extract latent topics within our corpus by including a set of covariates associated with the document collection of interest. STM encompasses document metadata as covariates into the prior distributions for document-topic proportions and topic-word distributions, including additional information. The model can be decomposed into three sub-models: (a) *a topical prevalence model*, which controls how documents are allocated to topics as a function of covariates; (b) *a topical content model*, which controls words frequency in each topic as a function of covariates; (c) *a core language model*, which combines the two sources of variation to produce the actual words in each document. Formally, documents are indexed as $d \in \{1....D\}$, while words are indexed as $n \in \{1…N\}$. The

observations $w_{d;n}$ are occurrences of words from a vocabulary indexed by $v \in \{1…V\}$. The number of topics is set by the researcher and indexed by $k \in \{1…K\}$. The document-level additional information is represented by two matrices. X is the matrix with topical prevalence covariates, while Y is the matrix with topical content covariates. The rows of these matrices - each representing a vector of covariates for a given document - are denoted by x_d and y_d, respectively**.** The process starts by drawing the document-level attention to each topic from a logistic-normal generalised linear model, based on a vector of covariates $\mathbf{x_d}$ and considering a *P x (K – 1)* matrix of coefficients *r* for the topical prevalence and a *(K-1) X (K-1)* $_d$, the topic-specific distribution over words is formed by representing each topic *k* with the V - dimensional baseline word distribution **m** (Airoldi et al. 2004), the topic specific deviation $K_k^{(t)}$, the covariate group deviation $K_{yd}^{(c)}$ and the interaction between the two just described. Finally, for each word in a document, topic assignment based on the document-specific distribution over topics and word assignment to a topic is drawn from multinomial models. The latter represents the core language model.

4. Data

We extracted status from Twitter using the open academy access API. We used the official campaign hashtag #DDLZAN. The volume of data was 123.268 tweets and 572.000 retweets. Given the large number of documents, we filtered the data. From the tweets' database we selected only the most active authors, who published more than 10 tweets in the stream. Moreover, we used the retweets database to select the influencers from the collected authors. We used the interactions between users to select the ones who were more central in the discourse. Briefly, we selected the most cited, quoted and mentioned authors. This interaction-based criterion allowed us to focus on the core of the data stream and study the influencer opinions. The number of influencers selected is 406. We selected these users from our database, and we aggregated their posts (Song et al. 2014). We calculated the interactivity measures for each author: amplification and endorsement. We have operationalized these concepts as quantitative variables. We operatized the amplification as the mean of the post's retweets and the endorsement as the mean of likes of the posts. We created a database with each author as record and published texts, amplification, and endorsements as fields. The posts are the content variable, and the interactivity measures the covariates associated.

5. Results

Twitter data are unstructured, so it's necessary to perform some phases of pre-processing for having structured data. There are different steps:

- Tokenized the documents, obtaining a set of distinct strings (tokens)

separated by spaces or punctuation marks;
- Normalized the text, so convert all the letters of the texts into a lower case;
- Removed special characters, punctuations, and numbers from the dataset. Also, special characters, hashtags (i.e. #ddlzan), symbols and stopwords are eliminated;
- Defined a grammatical tagging, which is the process of marking a word in a text as corresponding to a particular part of speech. In this case, we considered the nouns, verbs, and adjectives.

The pre-processing phase returned a database composed by 17353 tokens, 1182 type and 71 documents. The documents terms matrix dimension was 1182x71. Once this is done, STM can be applied to identify latent themes within the textual corpus. To define the optimal number of *k* topics to extract, we considered the measures of held-out likelihood and semantic coherence (Fig.1). Generally, topic models with excellent fit turn out to be higher on measures of held-out likelihood and semantic coherence (Roberts et al., 2016). So, we selected $k = 5$.

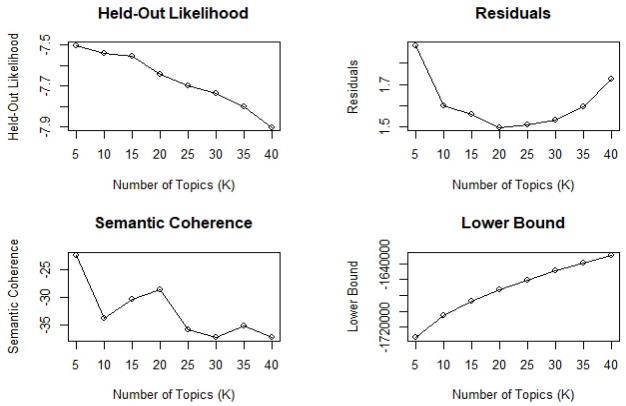


Figure 1. Diagnostic values by number of Topics

In fig.2 there are the top 10 frequent terms associated with the topics. We defined the first topic as "Law principles" because the terms associated with it allow to understand the content of DDLZan, including the fight against "homophobia" and violence motivated by "sexual" orientation and "gender identity". The second topic concerns the "Propaganda of the right-wing" against the approval of the bill, which has tried to scuttle the law by delaying the vote and asking for continuous changes to DDLZan. In particular, the contrary positions expressed by opposing politicians reiterated concerns already expressed by the Vatican.

Following this, the third topic concerns the criticism made by the center-left parties to Matteo Renzi and his party Italia Viva that, even though they are part of the government majority, have expressed perplexity about the contents of the proposal, asking for an agreement that would allow the law to be approved without including the protection of gender identity. Topic 4 highlights the demands of the majority parties to get to a vote soon, due to continued delays and attempts at ostracism by the opposition. In addition to this, the last topic concerns the democratic and public participation of citizens, who wish to have a law that allows protecting minorities comparing with other European countries.

Table 1. Top 10 frequent terms associated with topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Libertà	Senato	Votare	Approvare	Parlare
Omofobia	Renzi	Senato	Diritto	Italiano
Genere	Affossare	Renzi	Chiedere	Pensare
Donna	Destra	Chiedere	Parlare	Sinistra
Gay	Diritto	Muro	Genere	Approvare
Sinistra	Votare	Numero	Votare	Scrivere
Identità	Voto	Politico	Identità	Opinione
Parlare	Politico	Approvare	Paese	Politico
LGBT	Vaticano	Segreto	Odiare	Discriminazione
Sesso	Chiedere	Accordo	Anno	Fedez

6. Conclusions

Twitter encourages the formation of both social proximity networks and interest-based ties that can lead to filter bubbles (Marichal et al., 2020). So, we believe that by capturing the most important stakeholders in the debate, we can better understand the dynamics of stakeholder engagement. This happens because more and more individuals, including politicians, actors, bloggers, express their opinions and their daily activities on platforms and influence users by allowing content to spread. Fig. 3 and Fig. 4 plot a Loess, smoothed line of contributions plotted with confidence intervals on the predicted topic proportions for the measure of endorsement and amplification. According to amplification, we visualized that topic 4 has a major value, referring to the need that Twitter users

expressed about having a law against homophobia, while the topic 5 that expressed the participation of citizens has a higher value of endorsement.

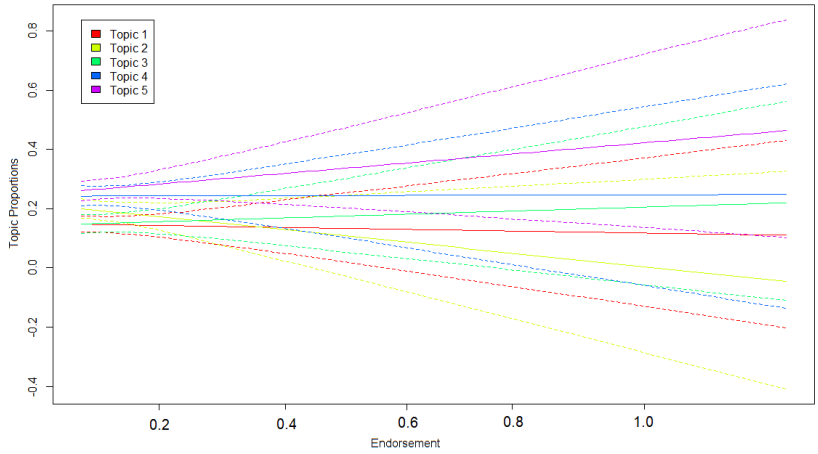


Figure 2. Topic proportions by endorsement

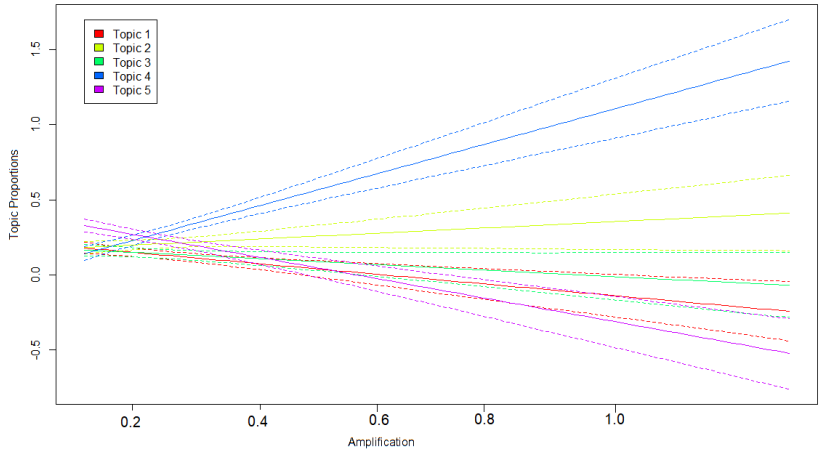


Figure 3. Topic proportions by amplification

References

Airoldi E. M., Cohen W. W. and Fienberg S. E. (2004). Bayesian models for frequent terms in text. In *Proceedings of the Classification Society of North America and INTERFACE Annual Meetings*

Blei D.M. and Lafferty J.D. (2007) A correlated topic model of science. *The Annals of Applied Statistics* 1(1):17-35
Castelló I. and Lopez-Berzosa D. (2021). Affects in Online Stakeholder Engagement: A Dissensus Perspective. *Business Ethics Quarterly,* 1-36.
Chatziadam P. Dimitriadis A., Gikas S., Logothetis I., Michalodimitrakis M., Neratzoulakis M. and Kondylakis H. (2020). TwiFly: A Data Analysis Framework for Twitter. *Information, 11(5)*, 247.
Colleoni E. Rozza A. and Arvidsson A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2), 317–332.
Gokulakrishnan B., Priyanthan P., Raghavan T., Prasath N., and Perera A., (2012) Opinion mining and sentiment analysis on a Twitter data stream, *International Conference on Advances in ICT for Emerging Regions,* 182-188
Grajzl P. and Murrell P. (2019), Toward understanding 17th century English culture: a structural topic model of Francis Bacon's ideas, J. *Comp. Econ.,* 47 (1)
He L., Han D., Zhou X. and Qu Z. (2020). The voice of drug consumers: online textual review analysis using structural topic model. *International Journal of Environmental Research and Public Health,* 17(10), 3648.
Kwon K. H., Chadha M. and Wang F. (2019). Proximity and networked news public: Structural topic modeling of global Twitter conversations about the 2017 Quebec Mosque shooting. *International Journal of Communication*, 13, 24.
Marichal J. and Neve R. (2019). Antagonistic bias: Developing a typology of agonistic talk on Twitter using gun control networks. *Online Information Review*.
Mishler A., Crabb E. S., Paletz S., Hefright B. and Golonka E. (2015). Using structural topic modeling to detect events and cluster Twitter users in the Ukrainian crisis. *In International conference on human-computer interaction* (pp. 639-644).
Roberts M. E., Stewart B. M. and Airoldi E. M., (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association 111(515)*, pp. 988–1003.
Sobkowicz P., Kaschesky M. and Guillaume B. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions on the web. *Government Information Quarterly*, 29(4), 470–479.
Song G., Ye Y., Du X., Huang X. and Bie S. (2014). Short text classification: A survey. Journal of multimedia, 9(5), 635.
Törnberg, A. and Törnberg, P. (2016). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media*, 13, 132-142.

From meta-data to research-data. How to use text analysis techniques to do Information retrieval as a propaedeutic practice to Data literacy

Deborah Grbac

Università Cattolica del Sacro Cuore di Milano –
deborah.grbac@unicatt.it

Abstract
Thanks to recent evolution of information technology and computational methods, text analysis is becoming a widespread tool adopted in different sectors and research disciplines, no more confined to computational linguistics. In particular 'distant reading', or computer-aided tools to retrieve information from texts, can be used in libraries too. Moreover, all the different phases and skills to be implemented to do a distant reading (collecting data, analysing data, retrieving relevant information from data, visualize results, finding out trends or patterns) have the further benefit to oblige to start to work with data, or in other words, text analysis is a good example to be used by librarians to start to offer some Data literacy courses to students.

Keywords: Information retrieval, text analysis, data literacy, research data

1. (Meta)data literacy at the library

Dealing with metadata, or data containing information about bibliographic references can be the first activity to introduce students to Data literacy, or all necessary knowledge and skills to do research using data, from a Data science perspective. In the past only software managing bibliographic data were at disposal and they were presented to students as a tool to automatically download citations. That software does not allow to work directly with metadata, which are only harvested and collected in a personal account to create a customized dataset reusable to draft bibliographies in different bibliographic (citation) styles. Instead, nowadays, statistical text analytics, using words frequencies, applied to metadata and texts to discover contents (Information retrieval) can also be used to gather metadata to build up datasets containing bibliographic references, selected according to a specific search query. A collection of metadata, to which further texts analytics can be added, can be used as a quantitative support for

scientific conclusions, or a statistically meaningful proof sustaining a theory; in a time when research is becoming more and more a data-driven one.
Corpora analytics, borrowed from linguists, such as stylometric based on ngrams calculation, normally used for authorship's attribution, can be used under the 'distant reading' approach to 'read' through statistical measures a whole database, made of references and corresponding full text, tackling it as a 'bag of words'. For distant reading we mean Franco Moretti's theory, nowadays a paradigm, or automatically reading great quantities of electronic texts (full-texts collections), through computational analyses producing statistical measures (Moretti, 2013). By 'distant reading' we can gather some ideas on the contents (preliminary analysis), but we are not yet properly analysing contents (contents are extracted instead by a *corpora* analysis). Looking at distance an entire database could be useful to do some information retrieval, if information needs strictly correspond to database contents, only in this case, words retrieved by statistical measures will in some ways resonate (semantic meaning) with effective content

1.1 Information retrieval vs. corpora analysis

Thanks to the big number of documents at disposal on electronic format nowadays, a researcher, who has selected a collection of texts, according to a stated query on a database, is able to retrieve some hints on contents, mainly on the statistical and linguistic point of view. Or through lexicometrics measures (words frequency, co-occurrence, ngrams, …) it is possible to infer trends and evolutions made explicit by customized visualisations which present data to support or disprove a theory or a model taken from the reality, if texts are coherent and meaningful.
Those analyses became *corpora* analysis, or a '*corpus*-based' approach, when computational assistance is applied to an assembled *corpus*, not only with statistically meaningfulness (as above), but a *corpus* build to answer a specific question and duly gathered for this purpose. Building a *corpus* allows to make a deeper analysis also according to contents, because *corpora* can be processed through for instance some cleaning techniques. The 'cleaning' allows to get rid of the 'rumours' and to make a better analysis in terms of words and their 'semantic' content. The difference between using statistic methods on texts to do a distant reading (a bird's eye approach) most of the time according to a query, or contents' analysis (meaning) done choosing every text to be part of a *corpus*, is that one has to go back and forth from statistical numbers to 'semantic' meaning of words and vice versa (Tuzzi, 2003).
The difference is not a minor one because when you apply 'distant reading' techniques to databases made of references and full texts, first of all contents are not picked up from a chosen *corpus*, but they are automatically extracted by search engines, and secondly contents are continuingly growing and diminishing.

Each month new periodicals' issues may be added to the database, or subtracted, according to the library subscriptions 'moving-wall' or agreed changing availability in contents during the time. These two conditions or the adding and subtracting of texts make the dataset produced by the query very changeable day after day and because of this different than a chosen *corpus*.
However, different databases are recently offering text analytics to browse contents by a query, creating a little bit of puzzle around on the theoretical and practical point of view. They are using both 'topic modelling' for texts classification and TF-IDF ratio to rank results. The application of these text analysis techniques can obtain some preliminary analysis or exploring ('discover') contents and 'scoring' ('relevance') results. If some contents can be inferred without reading texts ('distant reading'), results still have to be validated, as it can also happen with the classic "exact match" approach ensured by index search done on database contents.[1]
Those instruments are offered with the idea of enabling to use some text analysis techniques embedded to the database, avoiding recurring to specific external software (as external "search engine", or other software doing texts analysis automatically) and implementing instead customized notebooks (to databases' contents and user's needs) hosting the coding necessary to do the statistical analysis on data directly browsed from the database.
Regardless of which data can be downloaded and which analysis can be done, all those instruments are examples of how to use data to do research or data that will produce further research data (publications, datasets, protocols). Although, under a reference perspective it is essential to help students to understand what those different instruments are offering.

1.2 Data literacy should take into consideration existing caveats

When doing information retrieval in an automatic way, it is also essential to verify some points, because the easiness in retrieving results must be counterbalanced by a critical approach. Some questions have to be taken into consideration. Which resources are used to do the analysis? Resources are at disposal because they are subscribed, open access? Which are the contents, databases are multi or mono disciplinary? Which is the time range coverage, are they archives or collections of current texts?
Which and if database contents can resonate to the query formulated by the user?

[1] Another example of database using text analysis techniques to suggest "key words" to users in exploring and discovering contents is the platform under construction GoTriple a multilingual and multicultural discovery solution for the social sciences and humanities. Retrievable at: https://www.gotriple.eu/.

Is the key words strategy sufficient to retrieve results? Are contents numerically enough to be statistically significant? Statistical measures are working only if there is a big quantity of data at disposal, is there enough data to make a sensible analysis? Or databases consulted are enough to give a picture of the state of the art in the topic?
Which instruments are used to do the text analysis? In the past textual analysis software packages offered an all-in-one approach to do the analysis, users had only to do the query, without knowing how the software was working. Nowadays this approach is no more feasible because if science has to be reproducible (Open Science), it is also necessary to know which analyses are done and the steps followed to come to a result (protocols, codebooks, …). This is part of the 'research integrity' essential to do data-driven and reproducible research.
It is also important to know exactly which are the limits of the analyses that can be done because this has consequences on the results, or if there is a 'moving-wall' in contents availability, if a good OCR work has been applied to texts, if there is the opportunity to customize the analysis according to needs, for instance, linguistic ones, or the necessity to use a particular dictionary of words, if entity recognition is present or it can be modified.
Another important point is the use of authoritative sources as taxonomies to help the analysis, or the importance of *thesauri* and ontologies, all instruments that request an investment in time and money, but in case of less huge *corpora*, that can help in some ways to get a better result, counterbalancing a lack of significant statistical meaningfulness.
Finally, but not least, how to interpret the results? Which is the meaning of visualisations proposed? Are they clear to be interpreted by everyone? How to communicate results having particular attention to open data, but also data obtained under licence (also a CC BY licence) or data at disposal against subscription? This is also an important point in an era when digital and data skills are still lacking (also on copyright and licence issues), but the offer of data is continuously growing.

2. Some examples of text analysis used to do Information Retrieval

We will present in the following two instruments showcasing what can be done with text analysis for information retrieval, both instruments offer users an interface and analyses are done automatically, only results are shown, but information on their working is retrievable. In 2017 Jstor Labs at Ithaka[2]

[2] Jstor Archive and Jstor Labs are part of Ithaka a not-for-profit organisation preserving scholarly records (Jstor Archive) and proposing innovative pedagogic and research methods (Jstor Labs). Jstor Archive, one of the most ancient datasets containing references and full texts, is retrievable at: https://www.jstor.org/, Jstor Labs are retrievable at: http://labs.jstor.org/.

introduced one of the first instruments to do information retrieval in all the Jstor Archive database using text analysis, the instrument is called "Text Analyzer".[3] Users can simply upload a document in any format and language and retrieve further reading suggestions from the database contents. The uploaded text is read by two different topic modelling techniques an unsupervised and a supervised one. The last one recurs to a *thesaurus* (a set of humans' tagged labels) helping to identify concepts, entity recognition is possible (named entities: persons, locations, organizations), finally, another instrument helps to manage data. The result is the extraction of a list of terms that the user can adjust according to personal information retrieval needs and further use to search on the database. Proposed readings suggestions are scored by the TF-IDF calculation and are retrieved according to the co-occurrence of reading suggestions' terms with the terms extracted from the uploaded text (Snyder R. (2017).

A couple of years later, end 2019, a similar instrument was proposed by DESA (Department of Economic and Social Affairs) of the United Nations General Secretariat in cooperation with the United Nations Dag Hammarskjöld Library in New York. The instrument is based on the linked data technology, but it also uses some text analysis techniques.[4] Text analysis techniques aim at identifying the presence of SDGs (Sustainable Development Goals)[5] inside texts by a semi-supervised topic modelling trained with a set of data,[6] whose analyses have been further implemented using three authoritative sources: the former UNBIS (United Nations Bibliographic Information System), the European Union's legislation *thesaurus* called EuroVoc, and the SDGs (Sustainable Development Goals) taxonomy created by DESA (Department of Economic and Social Affairs).[7]

The DESA's taxonomy is also the interface to query the SDGs' ontology called SDGIO (Sustainable Development Goals Interface), whose aim is to draft a machine-readable representation of human knowledge in the SDGs' domain, to guide information retrieval, data extraction and analysis.[8]

[3] The "Text Analyser" is retrievable at: https://www.jstor.org/analyze/.

[4] The "Linked data service" is retrievable at http://metadata.un.org/?lang=en.

[5] Sustainable Development Goals, or SDGs, are 17 targets that have been adopted by United Nations Member States in 2015 to be realized by 2030. More information on SDGs is retrievable at: https://sdgs.un.org/goals.

[6] The first Topic modelling experiment was done on a set of 267 individual DESA publications. (LaFleur, 2019).

[7] The UNBIS thesaurus, now substituted by the United Nations Digital Library, is managed, and maintained by the Dag Hammarskjöld Library and retrievable at: http://metadata.un.org/thesaurus/?lang=en. The EuroVoc thesaurus is managed by the European Union Publications Office and retrievable at: https://op.europa.eu/it/web/eu-vocabularies.

[8] The ontology has been created and it is maintained by the UNEP's (United Nations Environment Programme) SDG and Environment Statistics Unit. (Coppens & Ghafary, 2018).

All those instruments have been applied to map and retrieve data and information on SDGs in the continuingly growing *corpus* consisting of all the texts produced by the United Nations. To have an idea about how it works a demo app is proposed called the "Linked SDG" service. When uploading a file or a web address, whose content is related to SDGs, the app uses controlled vocabularies, to extracts geographic entities, main concepts related to sustainable development and SDGs present in the uploaded text and it clusters them according to the 17 SDGs goals and sub-goals.[9]

2.1. Some examples of open research data and open education

At the beginning of 2022, after two years spent on working on the prototype, Jstor Labs launched the Constellate service, a platform for learning and performing text analysis, building datasets, and sharing open education analytics course material.[10] When accessing to the platform everyone is free to build one's dataset up to 25,000 units (ten dataset per day). Datasets consist of bibliographic data (metadata), language statistics measures (ngrams) and full texts available in open access (after request). Data visualizations are also at disposal: presentation of world frequencies in chronological order (time series), subject categories distribution in time, categories' tree-map, words bubble for terms frequencies. At least one visualisation is free for everyone.

Besides metadata and linguistic measures (ngrams, bigrams and trigrams) downloadable for free as open research data, the most interesting feature consists in the set of open education Jupyter notebooks written to have some clear and explained command line codebooks to do some text analytics autonomously. Proposed features are extracting bibliographic references (metadata), calculating world frequencies, scoring words (TF-IDF), discovering contents (topic modelling).[11] The notebooks are a set of instructions for a beginner in Digital Humanities downloadable from GitHub free for all, but also a computational space where to upload personal notebooks for more advanced users, this feature under subscription.

The Pierre Auger Observatory Open Data release.[12] Launched in October 2021 the initiative is another example of data education and open research data, the

[9] The "LinkedSGD.org" demo app is retrievable at: http://linkedsdg.apps.officialstatistics.org/#/.

[10] The Constellate platform is retrievable at: https://constellate.org/.

[11] Notebooks are set by default to work with 1.500 item taken from the dataset, but it is possible to broaden the range to work on the full length of the dataset available, or 25.000 items for free and 50.000 items under subscription.

[12] The Pierre Auger Observatory's website:
https://opendata.auger.org/#:~:text=The%20Pierre%20Auger%202021%20Open,Auger%20Collaboration%20Open%20Data%20Policy.

release is a tool enabling users to download data and to access to open education Jupyter notebooks helping users to do some autonomous analyses with those data. This is possible by a dedicated website hosting a framework to help people to work with data with all explanations and a set of simple analyses on Python notebook (a couple of tutorials on how to use data and five notebooks as an example reproducing results).[13]

3. Data literacy for librarians, but not only

At the beginning of this paper, we have talked about how Data literacy can help Library users, but how librarians could be trained to offer Data literacy to users? Or which is the Data literacy for librarians?

In the past an initiative called "Library Carpentry" helped librarians to start to learn how to use computational instruments and methods as: coding, organizing datasets, analyse data. At first to learn contents one has to enrol, but since 2019 course materials have been made open to all under a CCBY licence.[14]

More recently, an initiative coming from librarians and addressed to librarians has been proposed by the Federal Reserve Bank of St. Louis,[15] Research Division "Data Literacy for Librarians".[16] An online course developed by the Library Services and Economic Education teams, offering a professional development program using the open research data service called FRED[17] to build capacity "to use data effectively and accurately".

Seven areas have been identified to develop competencies on Data literacy or in general to find and reuse data. They are Information retrieval, authoritative sources, measures and the methodology, visualization of quantitative data, communicate data, make decision with data and ethics.

All the above-mentioned areas of data literacy are part of the 'research data life cycle'. If we deal with data, irrespective of the scientific discipline (linguistics, economics, or astronomy), the process is always the same: finding, collecting, analysis, communicate, take decision with data. Therefore (meta)data literacy is not only propaedeutic to research data, but it is part of it, and it is a knowledge that has to be known by everyone, not only librarians.

[13] Data at disposal are: 10% of the Observatory's cosmic-ray data and 100% of weather and space data.

[14] The Library Carpentry initiative website: https://librarycarpentry.org/lessons/.

[15] One of the 12 regional Federal Reserve Banks part of the decentralized central bank system having reserves and decision making distributed across the United States.

[16] The initiative is retrievable at: https://research.stlouisfed.org/info-services/data-literacy-for-librarians/.

[17] FRED, Federal Reserve Economic Data is a free online database containing economic data time series and tools to help users "to understand, interact with, display and disseminate data", with annotations helping to "understand and interpret data". Retrievable at: https://fred.stlouisfed.org/#.

References

Coppens L. and Ghafary D. (2018) *Sustainable Developments Goals Interface Ontology (SDGIO). Progress.* 24th September 2018, retrievable at: https://unstats.un.org/unsd/unsystem/Documents-Sept2018/Presentation-SDGIdentifiers-UNEP.pdf

LaFleur M. T. (2019). *Art is long, life is short: an SDG Classification System for DESA publications*. DESA Working Paper No. 159, United Nations Symbol ST/ESA/2019/DWP/159

Moretti F. (2013). *Distant reading*: London, New York, Verso

Snyder R. (2017). *Under the Hood of Text Analyzer*. Published on 7th March 2017, retrievable at: https://labs.jstor.org/blog/under-the-hood-of-text-analyzer-2/

Tuzzi A. (2003). *L'analisi del contenuto: introduzione ai metodi e alle tecniche di ricerca*, Roma, Carocci.

A comparison between machine learning and non-supervised techniques using the GP Index

Francesca Greco, Raffaella Gallo, Alessandro Polli, Fiorenza Deriu

Sapienza University of Rome – francesca.greco@uniroma1.it
raffaella.gallo@uniroma1.it alessandro.polli@uniroma1.it
fiorenza.deriu@uniroma1.it

Abstract
This paper aims to present a methodology for monitoring and assessing the public perception on social media. The assessment is based on a procedure using sentiment analysis to build a composite indicator, the Greco-Polli (GP) index. To this aim, we identify a set of hashtags according to the agenda setting to collect data from the Twitter repository and build a corpus for each hashtag. The procedure operates in two steps, and it is applied to the set of corpora, one for each hashtag. First, we automatically assess the sentiment using a non-supervised sentiment analysis procedure based on Emotional Text Mining (ETM), and we used the dataset already labeled by ETM to classify the sentiment by Machine Learning (ML). Then, we calculate the GP index for each sentiment analysis procedure, i.e., we transform the results in a stream of numerical data, and we build a composite index aggregating the information on the sentiment related to topics. We assess the applicability of the automatic labeling system by ETM for the training of ML algorithms and we compare the results by the GP index between ETM and ML. This procedure is applied to a case study on the public perception of security. Results highlight that this approach enables the assessment any potential change in the public perception, and that the use of ETM procedure for the automate labeling and the training of ML algorithms produces good results.

Keywords: Emotional Text Mining, Machine Learning, public perception, social media

1. Introduction

Social platforms are playing a growing role in our lives. We are constantly exposed, passively, or actively, to a flow of information that affects our lives and well-being especially in the last two years of the pandemic in which social interactions decreased to the benefit of the virtual ones. The shift from traditional media to social platforms, and their role in affecting people's perception and well-being, bring both risks and opportunities (Greco and Polli, 2021). The

seemingly bidirectionality of communication encourages the expression of opinions and feelings. Access this kind of information could be useful in several field such as election campaign (Fronzetti Colladon, 2021), public health (Boccia Artieri et al., 2021), education (Misuraca et al., 2021) and marketing (Gloor et al, 2020; Greco and Polli, 2020).

We refer to procedures aiming to extract such information from text as sentiment analysis or opinion mining (Liu, 2012). These are several statistical approaches ranging from unsupervised procedures, such as Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020), to supervised ones, such as machine learning (e.g., Ceron et al., 2016). These procedures measure opinions, beliefs, or personal judgments which are not based on rational reasoning, but on subjective emotions and, for this reason, they are suitable methods to study social media communication. Given the great relevance of security in determining people's well-being, this paper aims to present a methodology for monitoring and assessing the public perception on Twitter. The assessment is based on a procedure using sentiment analysis to build a composite indicator of the public perception.

In this study we present a methodology for monitoring and assessing the public perception on social media using a two step procedure. First, we automatically classify the sentiment using a non-supervised sentiment analysis procedure, ETM, and we use the ETM results for the automate labeling and the training of Machine Learning (ML) algorithms. Then, we calculate the GP index for each sentiment analysis procedure, transforming the results in a stream of numerical data, and we build a composite index aggregating the information on the sentiment related to topics. Finally, we assess the applicability of the automatic labeling system by ETM for the training of ML algorithms and we compare the results by the GP index between ETM and ML.

2. Measuring public perception by a composite indicator

In a previous study (Greco and Polli, 2021), the public perception of security was assessed by means of an unsupervised sentiment analysis procedure, Emotional Text Mining (ETM) and a composite indicator, the Greco-Polli (GP) index. To this aim, we identified three hashtags according to the agenda setting to collect data from the Twitter repository obtaining three corpora, one for each hashtag.

In particular, we scraped all the messages from the Twitter repository, written in Italian, containing the hashtag "security", "migration" and "crime" for two weeks, from April 5th to April 18th, 2019. collecting a total amount of almost 125 thousand messages (table 1).

Table 1. Corpora characteristics (Greco and Polli, 2021)

	Criminality	Migration	Security
Tweet	21,903	50,692	51,725
Retweet %	72.3	72.7	66.6
Token	432,807	956,002	1,000,264
Type	17,972	29,893	40,307
Hapax	9,387	16,721	22,656
TTR	0.04	0.03	0.04
Hapax %	52.2	55.9	56.2
N. terms in analysis	322	666	485
Tweet classified %	75.7	89.8	81.4

Then, we built three corpora and we measured the sentiment for each corpus by means of ETM, which is a social profiling method based on a multivariate analysis, i.e. a cluster analysis with a bisecting *k*- means algorithm on the term-document matrix to identify the topic representations and sentiments, and a correspondence analysis on the term-cluster matrix to detect the cultural symbolic categories setting the communication. The sentiment is calculated according to the representation, that could be qualitatively positive (+1), neutral (0), or negative (-1), and a numerical value was attributed to each tweet. Finally, we calculated the GP index to summarize the results related to the three hashtags, reducing them to a single stream of numerical data. The GP index embedded a penalty function, which reduced the weight of the sequences showing the greatest volatility over time.

The GP index aims to control the volatility of the tweet hourly sequences, in order to evaluate the public perception net of the potential presence of anomalous peaks that could hide the underlying trend in public perception. It is based on a double weighting scheme: the first weighting scheme relies on relative frequencies, which assign a weight directly proportional to the tweets production over time; the second one assigns a lower weight, in the composite indicator calculation, to the series that show greater volatility.

First, for each series, we assigned a conventional score (+1, 0, -1) to each tweet according to their positive, neutral and negative sentiment assessed by ETM. Then, we distributed the flow of tweets into hourly intervals, and we summarized the sentiment calculating the simple mean on an hourly basis. The main advantage of this procedure is that it replaces the stream of discrete scores with a stream of real values defined on the interval [-1, +1]. After calculating the frequency distribution of each series, it is straightforward to compute the related descriptive statistics, in particular the weighted mean and the standard deviation. In the second step, we aggregate the 3 sequences in a composite index using a

weighting vector. The weights are chosen to be inversely proportional to the standard deviation of the corresponding series. That is, the more a series shows volatility, the less it contributes to the computation of the composite index, neutralizing the effects due to the appearance of communicative hypes.

3. Method and main results

The machine learning procedure identifies the manifest sentiment expressed by Twitter users, while ETM identifies both the topic's representations and its sentiment. In the second stage, with the aim to build the GP index for each sentiment analysis procedure, we classify tweets by supervised Machine Learning models using ETM results to automatically label the tweet's sentiment. To apply supervised Machine Learning (ML) techniques, it is necessary to have a large number of labeled tweets. The labeling operation is inevitably very demanding because it requires a lot of time and resources. For this step of the analysis, therefore, we used the dataset already labeled by ETM, trying to pursue two goals: 1. to verify the applicability of the automatic labeling system by ETM for the training of ML algorithms; 2. to compare the results obtained by calculating the GP index on both the datasets labeled by ETM, and those labeled by ML models.

According to Liu (2012), to tackle with a text classification problem, we have considered each Tweet as a basic information unit (Liu 2012). This choice enabled us to apply different ML algorithms to the three datasets to identify those obtaining the best prediction results.

For a better performance of the ML models, the *corpora* were preprocessed removing stopwords, number, sparse terms as well as doing word stemming. For the extraction of the features, the document-term matrix was created according to the calculation of the TF-IDF scores.

A training set of 80%, and a test set of 20% of the tweets were created for each one of the three datasets. The same proportions – about 20% for the test set and about 80% for the train set – are respected approximately in relation to the specific categories, with the exception of the "neutral" categories for the "criminality" dataset and the "positive" category for the "security" dataset (tables 2-3-4). The variable with the sentiment score (+1, 0, -1), used as a prediction variable in ML models, has been recoded with nominal labels to include the category of unclassifiable tweets (unclassifiable, negative, neutral, positive).

Table 2. Criminality

Sentiment label	Test (%)	Train (%)	overall dataset (%)
unclassifiable	22,46	77,54	100
negative	20,94	79,06	100
neutral	5,95	94,05	100
positive	21,31	78,69	100
Total	**20**	**80**	**100**

Table 3. Migration

Sentiment label	Test (%)	Train (%)	overall dataset (%)
unclassifiable	22,63	77,37	100
negative	20,88	79,12	100
positive	14,40	85,60	100
Total	**20**	**80**	**100**

Table 4. Security

Sentiment label	Test (%)	Train (%)	overall dataset (%)
unclassifiable	23,90	76,10	100
negative	24,37	75,63	100
neutral	28,46	71,54	100
positive	8,33	91,67	100
Total	**20**	**80**	**100**

Since we used multi-class prediction models (the variable y is not dicotomic), weighted-average of precision, recall and F1-score were calculated, considering different number of units belonging to the various classes.
After comparing the performance of various algorithms, we chose the Support Vector Machines (SVM) and Random Forest (RF) models that achieved the best prediction results[1].

Table 3. ML Algorithms performance

	Criminality			
	weighted-average Precision	**weighted-average Recall**	**weighted-average F1-score**	**Accuracy**
SVM	0,846128738	0,841958457	0,838089477	0,84
RF	0,843437571	0,832816709	0,836411778	0,83
	Migration			
	weighted-average Precision	**weighted-average Recall**	**weighted-average F1-score**	**Accuracy**
SVM	0,797361413	0,791985599	0,791985599	0,79
RF	0,79169856	0,783836062	0,783355691	0,77
	Security			
	weighted-average Precision	**weighted-average Recall**	**weighted-average F1-score**	**Accuracy**
SVM	0,794559691	0,795923635	0,790241663	0,79
RF	0,78281392	0,783536974	0,783175447	0,78

[1] It is used RTextTools, a supervised learning R package for textual data classification (Jurka et al 2013)

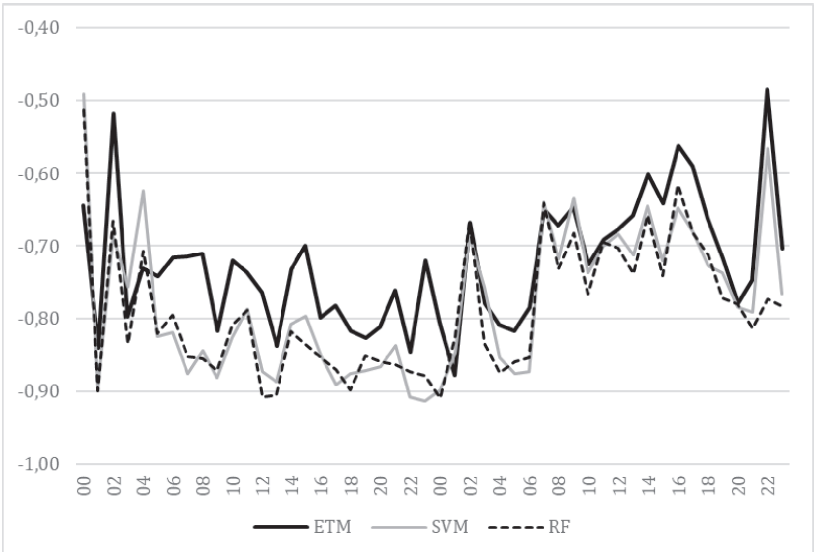
Encouraged by the good predictive performance of the models, in the next stage, starting from the labels predicted and assigned by the SVM and Random Forest models, we built the GP index for 20% of the total tweets (those included in the test sets).

The exam of Figure 1 allows us to highlight a substantial similarity between the three hourly streams, especially as regards the turning points. The main differences between ETM and ML methods occur on the first day when the volatility of the flows obtained with the ML methods appears less pronounced.

The correlation matrix shows a good agreement between the ETM and the SVM flows (ρ=0.79) and between ETM and RF ones (ρ=0.75). Moreover, the flows obtained through SVM and RF are strongly related (ρ=0.92).

However, a more in-depth analysis highlighted room for further improvements. We carried out a t-test at a significance level of 5% on the equality of the means of the three flows. The test rejects the null hypothesis of equality between means for the ETM versus SVM (p = value 0.007) and the ETM versus RF (p = 0.001), while it accepts it for the SVM versus RF (p = 0.548). We obtain the same result by conducting a Mann-Whitney test at a significance level of 5%. The test confirms a significant difference between ETM and SVM and between ETM and RF, while the test confirms the null hypothesis of no significant difference between SVM and RF.

Figure 1. GPI hourly stream for the three methods of machine learning



5. Conclusion

Results highlight that this approach enables the assessment any potential change in the public perception. The use of ETM procedure for the automate labeling and the training of ML algorithms produces good results even though the comparison of the GP indexes indicate that an improvement could be suitable.
In future studies it would be worth addressing the following further issues to improve the classification capacity: a) the differences in the corpus pre-treatment and terms selection between unsupervised and supervised procedures: b) the inclusion of unclassifiable messages in the training set that may interfere with both the training of ML model; c) and, finally, the sampling of the ML test tweets.

References

Boccia Artieri G., Greco F. and La Rocca G. (2021). Lockdown and Breakdown in Italians' Reactions on Twitter during the First Phase of Covid-19. *PACO, 14*(1): 261-282.

Ceron A., Curini L. and Iacus S. M. (2016). First- and second-level agenda setting in the twittersphere: An application to the italian political debate. *Journal of Information Technology and Politics, 13*(2), 159-174.

Fronzetti Colladon A. (2020). Forecasting election results by studying brand importance in online news. *International Journal of Forecasting, 36*(2) : 414-427.

Gloor P., Fronzetti Colladon A., de Oliveira J. M. and Rovelli P. (2020). Put your money where your mouth is: Using deep learning to identify consumer tribes from word usage. *International Journal of Information Management, 51,* doi:10.1016/j.ijinfomgt.2019.03.011

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F. and Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management,* 51: 101934, doi:10.1016/j.ijinfomgt.2019.04.007.

Greco F. and Polli A. (2021). Security Perception and People Well-Being. *Social Indicator Research,* 153(2): 741-758.

Jurka T. P., Collingwood L, Boydstun A. E., Grossman E., van Atteveldt W. (2013). RTextTools: A Supervised Learning Package for Text Classification. *The R Journal, Vol. 5/1*.

Liu B. (2012). *Sentiment analysis: Mining opinions, sentiments, and emotions. Sentiment analysis: Mining opinions, sentiments, and emotions.* Morgan and Claypool1.

Misuraca M., Scepi G. and Spano M. (2021). Using opinion mining as an educational analytic: An integrated strategy for the analysis of students' feedback. *Studies in Educational Evaluation, 68,* doi:10.1016/j.stueduc.2021.10097

Sentiment analysis dei corpora multilingua: La percezione pubblica del Covid-19 in Europa

Francesca Greco[1], Gevisa La Rocca[2], Giovanni Boccia Artieri[3]

[1]Sapienza University of Rome – francesca.greco@uniroma1.it
[2]Kore University of Enna – gevisa.larocca@unikore.it
[3]Carlo Bo University of Urbino – giovanni.bocciaartieri@uniurb.it

Abstract
This study investigates and compares the public perception of Covid-19 in Europe by means of the analysis of the public debate on Twitter with the hashtag #Covid-19 in the first phase of the pandemic. To this aim, over three million tweets were collected in Italian, Spanish, French, and German on March 4-11, 2020. These are the first four European countries that have been affected by the virus and introduced containment measures such as the lockdown. Although it was not possible to provide a geolocation of all the messages (less than 1% of the data), we chose a multilingual corpora analysis strategy, since the four corpora are the expression of a specific language spoken by several communities. Each language can be considered as a system in itself characterized by its semiotic and semantic aspects, which are difficult to translate. Nonetheless, they need to be considered to investigate the public perception since they reveal the complex symbolic-cultural system conveyed by the language. Accordingly, we use a text mining approach not needing a sentiment dictionary: Emotional Text Mining, which is based on a socio-constructivist approach and a psychodynamic model to detect the semiotic and semantic aspects conveyed by the tweets. The interpretation of the results made it possible to detect the symbolic-cultural matrix, the representations and sentiment of Covid-19 in the four languages and to compare them for the identification of the similarities and differences in the public perception during the first phase of the lockdown.

Keywords: #Covid-19, social media, Emotional Text Mining (ETM), public perception, corpora multilanguage.

Riassunto
Il presente studio indaga e confronta la percezione pubblica in Europa del Covid-19 attraverso l'analisi del dibattito pubblico attivato su Twitter e veicolato mediante l'hashtag #Covid-19 nella prima fase della pandemia in quattro paesi. A questo scopo sono stati raccolti oltre tre milioni di tweets tra il 4-11 marzo 2020 in italiano, spagnolo, francese e tedesco e che contenevano #Covid-19. Si tratta dei primi quattro paesi europei

che sono stati colpiti dal virus e che hanno dovuto introdurre misure di contenimento come il lockdown. Sebbene non sia stato possibile fornire una geolocalizzazione di tutti i messaggi (inferiore al 1% dei dati), si è scelto comunque di adottare una strategia di analisi per corpora multilingua, perché i quattro corpus sono espressione di un linguaggio specifico parlato da più comunità. Ciascun linguaggio si può considerare come un sistema a sé che si caratterizza per i propri aspetti semiotici e semantici difficilmente traducibili, un aspetto da tenere in considerazione se si desidera indagare la percezione pubblica, perché essa restituisce e rileva, in maniera importante, del complesso sistema simbolico-culturale connesso con il linguaggio. A questo scopo si è utilizzata la procedura dell'Emotional Text Mining (ETM), che consente di indagare gli aspetti semiotici e semantici caratterizzanti il testo a partire da un approccio socio-costruttivista di matrice psicodinamica, che non necessita di dizionari precodificati del sentiment e che identifica gli aspetti percettivi a partire dall'associazione delle parole. L'interpretazione dei risultati ha consentito di identificare la matrice simbolico-culturale, le rappresentazioni e il sentiment del Covid-19 nelle quattro lingue selezionate e di confrontarli per individuare le similitudini e le differenze nella percezione pubblica della prima fase del lockdown.

Parole chiave: #Covid-19, social media, Emotional Text Mining (ETM), percezione pubblica, corpora multilingua.

1. Introduzione

La diffusione di internet ha consentito quotidianamente a milioni di persone di navigare sul web, creare account e cercare o condividere informazioni, determinando un incremento nell'uso delle piattaforme digitali, come Twitter, Facebook, blog e microblog, che sono diventate delle fonti essenziali per conoscere le opinioni, i sentimenti e le emozioni. Per questo motivo, lo sviluppo della ricerca sui social media è stato associato alla rivoluzione dei big data. Per dare un'idea della quantità di dati prodotti, all'inizio del 2017 Twitter contava oltre 300 milioni di utenti che ogni giorno producevano una media di circa 500 milioni di messaggi in diverse lingue. Tuttavia, come evidenziato da King (2016), la sfida posta da questa rivoluzione non riguarda i dati, poiché oggi spesso è facile ed economico ottenerli, ma piuttosto l'estrazione di informazioni dai big data; vale a dire, quale metodo utilizzare per trasformare i dati in conoscenza, poiché essi sono costantemente prodotti, sono eterogenei e non strutturati (Feldman and Sanger, 2007).

La sentiment analysis, anche detta opinion mining, è un insieme di procedure supervisionate e non supervisionate di text mining volte a identificare e misurare le opinioni, le emozioni e i sentimenti. Queste tecniche hanno avuto un importante sviluppo negli ultimi venti anni, come testimonia la pubblicazione di circa ventimila articoli su questo argomento, di cui il 57% è costituito da atti di conferenze, a dimostrazione dell'innovatività e della continua crescita di questo settore di ricerca. Lo sviluppo e il successo negli ultimi due decenni della

sentiment analysis sono dovuti alla sua applicazione in quasi tutti i settori aziendali e sociali, in quanto le opinioni, determinando il comportamento, influenzano molte delle attività umane. Il suo successo, quindi, coincide con la sua applicabilità all'interno dei social media dove essa è diventata la tecnica di analisi più utilizzata, poiché si presenta come uno strumento veloce ed economico per studiare i fenomeni sociali.

In particolare, questo studio si è occupato di rilevare la percezione pubblica in Europa del Covid-19 attraverso l'analisi del dibattito attivato su Twitter e veicolato mediante l'hashtag #Covid-19 nella prima fase della pandemia. All'inizio del lockdown, in una sola settimana, sono stati prodotti oltre tre milioni di tweet in italiano, spagnolo, francese e tedesco a testimoniare l'intensa reazione emotiva alla pandemia. Si tratta, infatti, dei primi quattro paesi europei che sono stati colpiti dal virus e che hanno dovuto introdurre misure di contenimento come il lockdown. Come evidenziato da Bruns and Stieglitz (2012), gli stati di emergenza producono un forte impatto sulla percezione pubblica che si declina attraverso un aumento della produzione di messaggi sulle piattaforme digitali.

Sebbene la geolocalizzazione sia presente in meno dell'1% dei casi, si è scelto comunque di adottare una strategia di analisi per corpora multilingua, perché i quattro corpus sono espressione di un linguaggio specifico parlato da più comunità e perché lo scopo della ricerca è stato quello di individuare e confrontare la percezione pubblica del Covid-19 in questi paesi. Ciò ha comportato la necessità di scegliere una procedura di sentiment analysis che consentisse di confrontare i diversi linguaggi. Ciascun linguaggio, infatti, si può considerare come un sistema a sé che si caratterizza per i propri aspetti semiotici e semantici difficilmente traducibili. Tale aspetto è tanto più da tenere in considerazione se si desidera indagare la percezione pubblica (Greco and Polli, 2021), perché essa restituisce e rileva, in maniera importante, il complesso sistema simbolico-culturale connesso con il linguaggio.

2. Analisi emozionale dei corpora multilingua

Esula da questa trattazione la discussione degli importanti contributi forniti dalla linguistica computazionale sul tema dei corpora multilingua (e.g., Chiari, 2012; Balbi and Misuraca, 2006) e, più in particolare, sulla sentiment analysis di corpora multilingua (Schulz et al., 2010; Bond et al., 2015). Tuttavia, ci sembra importante sottolineare come la scelta della metodologia impiegata in questo studio è connessa con la difficoltà di reperire delle risorse affidabili e open-source per la misurazione del sentiment nelle lingue da noi analizzate.

L'analisi, quindi, è stata condotta con una procedura per la sentiment analysis, che non necessità di risorse per la classificazione e che consente di indagare sia il piano semantico che quello semiotico della comunicazione, favorendo in tal modo il confronto tra i diversi sistemi simbolico culturali. A questo scopo si è

scelto di utilizzare il metodo dell'Emotional Text mining (ETM – Greco, 2016; Greco and Polli, 2020). L'ETM, è una procedura non supervisionata che si basa su un approccio socio-costruttivista ad orientamento psicodinamico che effettua una profilazione sociale a partire dai testi e che è stata applicata in diversi ambiti (e.g., Cordella et al., 2020; Boccia Artieri et al., 2021°; 2021b; Di Trani et al., 2021; Greco and Polli, 2021; Passarello et al., 2021).
Dal repository di Twitter sono stati raccolti tutti messaggi che contenevano #Covid-19 pubblicati tra il 4 e l'11 marzo 2020 in italiano, spagnolo, francese e tedesco (tabella 1).

Tabella 1. Caratteristiche descrittive dei dataset dal 4 all'11 marzo 2020

	Italiano	*Spagnolo*	*Francese*	*Tedesco*
Retweets e @mentions	542.322 (78,7%)	1.050.817 (82,0%)	722.108 (87,9%)	159.690 (71,9%)
Tweet geolocalizzati	5.852 (0,9%)	2.791 (0,7%)	1.658 (0,3%)	1.206 (0,5%)
Tweet (Totale)	689.514	1.280.732	821.914	222.046
Documenti in analisi				
Type	103.601	145.343	76.089	81.514
Token	18.760.987	35.235.705	24.504.338	5.504.245
Type Token Ratio (TTR)	0,006	0,004	0,003	0,015
Hapax	35.608	47.849	23.350	30.395
Hapax%	34,4%	32,9%	30,7%	37,3%

Dopo una prima fase di pulizia dei dati, l'ETM prevede una sequenza di fasi di sintesi, dalla lemmatizzazione e selezione dei termini in analisi sulla base di indici lessicometrici (Giuliano and La Rocca, 2010; Greco, 2016), all'analisi dei cluster sulla matrice termini-documenti e quella delle corrispondenze sulla matrice termini-cluster. In particolare, dopo aver rimosso le parole vuote, vengono selezionati i termini del rango alto e medio di frequenza (Bolasco, 1999; Greco, 2016) e sulla matrice termini-documenti viene effettuata un'analisi dei cluster con algoritmo bisecting *k*-means utilizzando la misura di distanza del coseno (Steinbach et al., 2000), limitata a 20 partizioni, con l'esclusione dei messaggi che non presentano almeno una co-occorrenza. La partizione ottimale viene scelta in funzione dell'intraclass correlation coefficient, il Calinski-Harabasz e il Davies-Bouldin. Successivamente, viene effettuata un'analisi delle corrispondenze (Lebart e Salem, 1994) sulla matrice termini-cluster per identificare gli assetti simbolici caratterizzanti la comunicazione (Cordella et al., 2014; Greco, 2016).
Dopo aver effettuato l'interpretazione dei risultati, la matrice simbolico-culturale, le rappresentazioni e il sentiment del Covid-19 nelle quattro lingue, i risultati sono stati confrontati per individuare le similitudini e le differenze nella percezione pubblica sviluppata durante la prima fase del lockdown.

3. Risultati

I risultati hanno evidenziato come vi siano cospicue similitudini in Europa sia sul piano semiotico che su quello semantico (tabella 2). L'Italia è stato il primo paese ad identificare un caso infetto e a scegliere di ricorrere al confinamento per fare fronte alla diffusione del contagio.

Tabella 2. Quadro sintetico dei risultati dell'ETM dei quattro corpora

	Cl	*Tw%*	*Etichetta*	*Sent.*	*Fattore 1*	*Fattore 2*	*Fattore 3*
					Contagio	*Reazione*	*Categorie*
Italiane	1	29.5	Confinamento forzato	Neg	Sentimento		Tempo
	2	23.3	lockdown Italiano	Neg	Ragione	Policy	
	3	24.0	Minaccia Straniera di Contagio	Neg	Sentimento		Spazio
	4	23.2	Minaccia Nazionale: Zona Rossa	Neg	Ragione	Health	
					Diffusione	*Crisi*	*Livello*
Spagnolo	1	11.7	Panico	Neg	Locale	Sanitaria	
	2	10.6	Lockdown Spagnolo	Neg			Macro
	3	12.4	Informazioni di Prevenzione	Neu		Sanitaria	
	4	11.0	Impatto Economico	Neg	Locale	Economica	Micro
	5	19.4	Diffusione Mondiale del virus	Neg	Globale		Micro
	6	14.4	Impatto Mondiale sulla Salute	Neg	Globale		Macro
	7	10.6	Strategie Indiv. di Prevenzione	Neu	Locale	Sanitaria	Micro
	8	9.8	Confinamento forzato dei Paesi	Neg	Locale		Macro
					Impatto	*Livello*	*Crisi*
Francese	1	14.3	Contagio Francese	Neg			Sanitaria
	2	13.0	Crisi Nazionale	Neg	Produttivo	Macro	
	3	10.8	Lockdown Italiano	Neg	Sociale	Macro	
	4	9.7	Psicosi Nazionale	Neg	Produttivo		Sanitaria
	5	18.6	Diffusione Mondiale del Virus	Neg	Sociale	Micro	Economica
	6	10.8	Lockdown Francese	Neg		Macro	Sanitaria
	7	11.0	Impatto Economico	Neg	Produttivo	Macro	Economica
	8	11.8	Strategie Indiv. di Prevenzione	Neu	Produttivo	Micro	
					Impatto	*Contagio*	*Reazione*
Tedesco	1	13.6	Reazione del Governo	Neg	Sanitario		Attivo
	2	10.1	Cancellazione Eventi Sportivi	Neg	Sociale		Attivo
	3	14.1	Lockdown Italiano	Neu	Sanitario	Ragione	Passivo
	4	13.2	Panico generale	Neg		Sentimento	Passivo
	5	12.8	Diffusione mondiale del virus	Neg	Sanitario		Passive
	6	11.4	Impatto sulla Salute	Neg	Sanitario		Attivo
	7	11.2	Impatto Economico	Pos		Sentimento	Passivo
	8	13.6	Cancellazione Eventi Sportivi	Neu	Sociale	Ragione	

Ciò ha profondamente influito sugli altri cittadini europei e sulla percezione pubblica della sicurezza, come testimonia la presenza della parola "Italia" ed "italiani" in tutte le lingue esaminate e del tema del lockdown italiano. Sebbene vi siano delle specificità, in tutte le lingue è presente la distinzione tra impatto sociale e quello istituzionale, politico e sanitario, e tra il livello macro e quello micro. Infine, le piattaforme digitali hanno svolto il ruolo di agorà, deputato alla

discussione pubblica, garantendo e permettendo agli individui di condividere le proprie opinioni e i sentimenti, costruendo nuove dimensioni semiotiche per dare senso alle proprie esperienze.

4. Discussione e conclusioni

Come evidenziato dai risultati, indubbiamente, in questa fase della pandemia, i social media hanno svolto un ruolo importante costituendosi quale agorà virtuale per la discussione pubblica attraverso la quale le persone hanno condiviso le proprie opinioni, i sentimenti e le esperienze. Queste sono state raccolte sotto uno specifico hashtag che si è costituito quale contenitore semiotico di senso. Le piatteforme digitali, infatti, hanno svolto un ruolo importante nel supportare la costruzione e la comunicazione di storie attorno al tema del Covid-19, consentendo alle persone di riorganizzare la propria esperienza traumatica.

Analizzare i tweet relativi alla pandemia in italiano, spagnolo, francese, tedesco pubblicati nel periodo immediatamente precedente al primo dei lockdown europei ha consentito di analizzare le emozioni vissute, le riflessioni condivise e la discussione attivate in quelle prime ore da parte degli individui/utenti che hanno visto il Covid-19 invadere la propria realtà quotidiana. Ci permette, inoltre, di capire come, a partire dai primi casi italiani e dal lockdown introdotto in Italia, gli altri tre Paesi europei abbiano iniziato a prepararsi all'impatto della prima ondata di pandemia.

Le quattro lingue – che rappresentano i quattro paesi – sono parlate anche al di fuori dei loro contesti nazionali, a causa di storie di migrazione o colonizzazione. In questo modo, oltre a fornirci informazioni sui quattro paesi europei, questa analisi ci offre l'opportunità di tracciare le dinamiche del dibattito pubblico su ciò che sta accadendo in Europa – dentro e fuori i suoi confini – in relazione alla diffusione di Covid-19 e le scelte politiche che vengono fatte dalle comunità linguistiche target in altri paesi non europei.

Infine, la nostra ricerca propone un nuovo approccio per l'analisi del sentimento dei corpora multilingua che consente di studiare la comunicazione sui social media. La scelta di studiare e confrontare i sistemi simbolico-culturali veicolati dalle lingue oggetto di studio attraverso una opportuna procedura di text mining, invece del linguaggio stesso, ha offerto l'innegabile vantaggio di confrontare lingue differenti senza necessitare di dizionari di riferimento.

Bibliografia

Balbi S. and Misuraca M. (2006). Rotated Canonical Correlation Analysis for Multilingual Corpora. In JM. Viprey, C. Condé, A. Lelu and M. Silberztein, editors, *Proceedings JADT 2006: 8es Journées internationales d'Analyse statistique des Données Textuelles*, April 19-21, 2006, Besançon, France, pp. 31-42. Lexicometrica.

Boccia Artieri G., Greco F. and La Rocca G. (2021). The flame of the Coronavirus in Italy. Looking for fear arousing appeal during the media hype period on Twitter. *International*

Review of Sociology, 31(*2*), 287-309.
Boccia Artieri G., Greco F. and La Rocca G. (2021). Lockdown and Breakdown in Italians' Reactions on Twitter during the First Phase of Covid-19. *PACO, 14*(1), 261-282.
Bolasco S. (1999). *Analisi multidimensionale dei dati: Metodi, strategie e criteri d'interpretazione.* Carocci.
Bond F., Da Costa L. M. and Lê T. A. (2015). IMI - A multilingual semantic annotation environment. In *Proceedings of System Demonstrations,* pp. 7-12.
Bruns A. and Stieglitz S. (2012). Quantitative approaches to comparing communication patterns on Twitter. *Journal of Technology in Human Services*, 30(3-4): 160-185.
Chiari I. (2012). Corpora e risorse linguistiche per l'italiano: Stato dell'arte, problemi e prospettive. I *Italienisch,* 34(2): 90-106.
Cordella B., Greco F., Meoli P., Palermo V. and Grasso M. (2020). Educational culture and job market: A text mining approach. In D.F. Iezzi, D. Mayaffre, M. Misuraca (eds.), Text Analytics. Advances and Challenges, Springer, Heidelberg. ISBN 978-3-030-52679-5
Cordella B., Greco F. and Raso A. (2014). Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In E. Née M. Daube M. Valette and S. Fleury, editors, *Proceedings JADT 2014, 12es Journées internationales d'Analyse Statistque des Données Textuelles, Paris, France, Juin 3-6, 2014,* pp. 173-184. Lexicometrica.
Di Trani M., Spoletini R., Renzi A. Greco, F. and Scaravelli G- (2021). Psychologists' representations about their professional functions within assisted reproductive multidisciplinary teams. *Professional Psychology: Research and Practice, http://dx.doi.org/10.1037/pro0000355*
Feldman R. and Sanger J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.
Giuliano L. and La Rocca G. (2010). *Analisi automatica e semi-automatica dei dati testuali: Strategie di ricerca e applicazioni* (vol. II). Led.
Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.
Greco F. and Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management,* 51: 101934,
Greco F. and Polli, A. (2021). Security Perception and People Well-Being. *Social Indicator Research,* 153(2): 741-758.
King, G. (2016). Preface: Big Data Is Not About The Data! In R. M. Alvarez, editors, *Computational Social Science: Discovery and Prediction* (pp. Cambridge University Press.
Lebart L. and Salem A. (1994). *Statistique Textuelle.* Dunod.
Passarello N., Greco F., Piras F., Cordella B., Lai C., Spalletta G., Mandolesi L. and Laricchiuta D. (2021). "No words for feelings": a multidimensional analysis of the alexithymia construct. *Psychology Hub, 39*(2), 41-50.
Schulz J. M., Womser-Hacker C. and Mandl T. (2010). Multilingual corpus development for opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010,* pp. 3409-3412.
Steinbach M., Karypis G. and Kumar V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 400, pp. 525–526. Boston.

Leveraging Deep Learning models to assess the temporal validity of Emotional Text Mining method

Francesca Greco, Alessandro Polli, Federico Siciliano

Sapienza Università di Roma – francesca.greco@uniroma1.it
alessandro.polli@uniroma1.it
federico.siciliano@uniroma1.it

Abstract
The paper present the training of a surrogate model to mimic the functioning of a text mining social profiling method, Emotional Text Mining, applied on the Italian tweets containing #Covid-19 during the first lockdown. This surrogate model is based on state-of-the-art Deep Learning models in the field of Natural Language Processing (NLP). These models feature an architecture called Transformer, which adopts a mechanism of self-attention that weighs the significance of each part of the input data. The model was trained on the tweets and factors of an initial time period to predict the polarisation of a tweet according to each factor found using the ETM technique. The model was then applied to subsequent periods. Using appropriate metrics, it was then possible to quantify the need to re-run the technique from scratch on the following time periods, while also being able to assess which factors still remained relevant over time.

Keywords: Natural Language Processing, Emotional Text Mining, Neural Network, Machine Learning, Covid-19, Social Media

Riassunto
Il lavoro presenta l'addestramento di un modello surrogato per imitare il funzionamento di un metodo per la profilazione sociale, l'Emotional Text Mining, applicata sui tweet italiani che contenevano #Covid-19 prodotti durante il primo lockdown. Questo modello surrogato è basato su modelli di Deep Learning all'avanguardia nel campo del Natural Language Processing (NLP). Questi modelli sono caratterizzati da un'architettura chiamata Transformer, che adotta un meccanismo di auto-attenzione che pesa l'importanza di ogni parte dei dati di input.
Il modello è stato addestrato sui tweet e i fattori di un primo periodo di tempo per prevedere la polarizzazione di un tweet in base a ciascun fattore trovato con la tecnica ETM. Il modello è stato poi applicato ai periodi successivi. Utilizzando metriche appropriate, è stato poi possibile quantificare la necessità di rieseguire la tecnica da zero sui periodi di tempo successivi, potendo anche valutare quali fattori sono rimasti rilevanti nel tempo.

Parole chiave: Elaborazione del linguaggio naturale, Emotional Text Mining, Rete Neurale, Machine Learning, Covid-19, Social Media

1. Introduzione

L'ampia diffusione di Internet e la nascita dei Social media, ha permesso alle persone di cercare e condividere informazione su base giornaliera e rapidamente. Questa grande quantità di materiale generato dagli utenti può anche essere consultato facilmente e a basso costo; per esempio, il famoso social media Twitter fornisce API che permettono ai ricercatori di accedere al loro enorme archivio di dati. Il contenuto dei tweets può quindi essere elaborato utilizzando una metodologia per la profilazione sociale come l'Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020) per identificare il sentiment degli utenti dei social media su un argomento. Poiché le opinioni, le emozioni ed i sentimenti mutano nel tempo, diventa necessario applicare il metodo a più riprese per diversi periodi di tempo. Sfortunatamente, però, questo processo è computazionalmente molto oneroso perché è richiesto applicarlo a grandi volumi di dati integralmente. In questo articolo, prendiamo spunto dalle recenti scoperte nel campo del Deep Learning per cercare di (i) stimare quanto sia valido il risultato della tecnica dell'ETM in periodi successivi a quello su cui è stata utilizzata e (ii) quando si rende necessario ripetere l'analisi per intero.

2. Lavori correlati

2.1. Emotional Text Mining

L'utilizzo di tecniche di text mining per misurare opinioni, sentimenti ed emozioni degli attori sociali è molto diffuso e si basa fondamentalmente su due approcci: quello semantico, in cui le parole vengono considerate solo per il loro significato esplicito, e quello semiotico, in cui le parole vengono considerate come simboli dei modi di pensare delle persone. Quest'ultimo offre il vantaggio non indifferente di connettere la comunicazione con il comportamento, effettuando di fatto una profilazione a partire dalla comunicazione. L'Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020) adotta un approccio semiotico al testo e si basa su un modello socio-costruttivista ad orientamento psicodinamico che considera le emozioni non come delle sensazioni (correlato biologico) ma come l'espressione del funzionamento inconscio e sociale della mente (Carli, 1990; Greco, 2016). In tal senso l'ETM si configura come un metodo per la profilazione sociale poiché identifica le categorie simbolico-culturali che organizzano il comportamento, le sensazioni, le aspettative, gli atteggiamenti e la comunicazione degli attori sociali. Di conseguenza, l'ETM non guarda al contenuto esplicito della comunicazione ma alla sua modalità, vale a dire quali parole vengono selezionate e come vengono associate.

Mentre il funzionamento mentale procede dal livello semiotico a quello

semantico nel generare il testo, la procedura statistica simula il processo inverso del funzionamento mentale, dal livello semantico a quello semiotico. Per questo motivo, l'ETM esegue una sequenza di procedure di sintesi, dal pretrattamento del testo e la selezione dei termini, all'analisi multivariata, per identificare il livello semiotico, partendo da quello semantico (Greco and Polli, 2020). È una procedura automatica, veloce e relativamente semplice, che può essere utilizzata per estrarre informazioni in merito a quelle che potremmo definire le opinioni emotivamente guidate. A differenza di altre tecniche di sentiment analysis non solo misura il sentiment ma fornisce anche informazioni qualitative sullo stesso.

2.2. Natural Language Processing

L'elaborazione del linguaggio naturale (NLP) vede l'unione dei campi della linguistica e dell'informatica, focalizzandosi su come programmare i computer per analizzare grandi quantità di dati in linguaggio naturale. Sono vari i task specifici che ricadono all'interno di questo obiettivo più generale, come quello di categorizzare dei documenti, controllare se una notizia sia falsa o capire che sentimento l'autore stia esprimendo. Negli ultimi anni, il campo del NLP ha visto l'emergere di metodi di apprendimento basati su reti neurali, in particolare quelli appartenenti alla categoria del Deep Learning. Questi modelli, costituiti da architetture grandi e complesse, fanno uso di grosse quantità di dati e hanno dimostrato di riuscire a raggiungere risultati all'avanguardia in molti compiti del linguaggio naturale.

I primi modelli di rete neurale applicati a NLP erano basati su strutture di tipo ricorsivo (Recurrent Neural Network - RNN); queste ricevono i dati in maniera sequenziale, elaborando i dati ricevuti ad ogni istante e mantenendo un aggregato che riassume l'informazione passata. Nonostante vari miglioramenti, le RNNs sono state superate nell'elaborazione del linguaggio con i Transformers. Questi modelli adottano quello che viene definito Meccanismo di Attenzione (Vaswani et al., 2017), riescono cioè a guardare la frase in ingresso nella sua interezza ponderando il significato di ogni parola della sequenza. Questa caratteristica, oltre a far raggiungere risultati notevoli, permette una maggiore parallelizzazione rispetto alle RNN e riduce quindi i tempi di allenamento. Questo ha aiutato lo sviluppo di sistemi pre-addestrati come BERT (Devlin et al., 2019), allenati su grandi dataset linguistici, quali Wikipedia, e che possono essere messi a punto per compiti specifici.

3. Metodi

La tecnica di Emotional Text Mining permette di effettuare una profilazione degli utenti dei media attraverso i loro post. Tuttavia, questa suddivisione in diversi gruppi di profili è dipendente dal tempo, poiché l'esecuzione della procedura su un altro periodo di tempo può portare a risultati diversi. La procedura è però

computazionalmente costosa, e ciò non permette di eseguirla agilmente per diversi periodi di tempo. Per superare questo ostacolo, abbiamo addestrato un modello surrogato, basato su Transformer, per imitare il funzionamento della tecnica ETM. Abbiamo inoltre valutato e impiegato diverse metriche per quantificare la necessità di eseguire la tecnica da zero sui periodi di tempo successivi.

3.1. Modello

Sono stati raccolti tutti i tweet in lingua italiana contenenti la parola "coronavirus" successivamente a tre importanti conferenze stampa del Capo di Governo durante la prima fase di lockdown per un totale di 1.884.670 messaggi costituiti dall'81% di retweet. I 6 periodi temporali, nell'anno 2020, corrispondono a: 4-11 Marzo, 24-31 Marzo, 10-14 Aprile, 30 Aprile - 9 Maggio, 17-26 Maggio, 3-12 Giugno. Il primo sample dei dati (numero di tweets = 80.519) è stato analizzato con ETM (Boccia Artieri et al., 2021a; 2021b), per poi essere usato come input della rete neurale. In breve, l'ETM esegue un algoritmo di clustering sulla matrice termini-documenti e un'analisi delle corrispondenze sulla matrice cluster-termini, producendo un insieme di fattori, ognuno caratterizzato da un insieme di termini i quali ne danno il significato, e un insieme di cluster, posizionati nello spazio dei fattori. Per ogni tweet abbiamo perciò un cluster di appartenenza e perciò i fattori più significativi per esso. Maggiori informazioni possono essere trovate sul relativo articolo (Boccia Artieri et al., 2021a; 2021b).

Per strutturare la nostra rete neurale, siamo partiti da un modello pre-allenato basato su BERT, Neuraly[1], basato su una versione di BERT per l'italiano, Italian BERT[2], e messo a punto per effettuare sentiment analysis su un dataset di tweet in italiano, un task simile al nostro. La prima parte del nostro modello perciò genera la rappresentazione vettoriale della frase ricevuta in ingresso. Abbiamo poi aggiunto un'altra porzione di rete neurale specifica per il nostro task. L'output, viene poi passato a tre strutture parallele ognuna responsabile della predizione di un singolo fattore. Per ogni fattore i 3 neuroni di output ci danno una probabilità, rispettivamente quella dell'input di essere più vicino al polo negativo, al polo positivo o specificatamente a nessuno dei due poli di quel fattore.

3.2. Metriche

Per capire quanto il modello funzioni adeguatamente in periodi diversi da quello su cui è stato allenato, abbiamo selezionato delle metriche ad-hoc. Le prime 2 sono applicate separatamente per ogni singolo fattore, considerando le probabilità di essere polarizzati negativamente, neutralmente o positivamente.

La prima metrica, chiamata Max-Value, consiste nel prendere per ogni tweet il

[1] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment
[2] https://huggingface.co/dbmdz/bert-base-italian-cased

massimo di questi 3 valori, per poi effettuare una media. In questa maniera, vogliamo valutare l'incertezza del modello: tanto più questa metrica tende al 100%, tanto più il modello sarà certo della sua predizione, tanto più tende al 33%, tanto più il modello sarà incerto tra i 3 output possibili.
La seconda metrica, chiamata Poles-Diff, calcola per ogni tweet la differenza, in valore assoluto, tra le probabilità di essere polarizzato positivamente e negativamente, per poi mediare i risultati. Vogliamo così valutare quanto il modello consideri i tweets polarizzati secondo ogni fattore. Più il valore tende al 100%, più l'opinione delle persone è polarizzata, mentre più tende allo 0% e meno quel fattore è in grado di esprimere l'emozione del pubblico.
L'ultima metrica, Four-Frequency, prende invece in considerazione la clusterizzazione dei tweets; per ogni tweet, viene usato il modello per predire una polarizzazione per ogni fattore, scegliendo quella con più alta probabilità. I tweet vengono quindi divisi in cluster in base ai valori ottenuti per ogni fattore, e selezioniamo i 4 cluster più presenti. Più i 4 valori si avvicinano al 25%, più l'opinione del pubblico è equamente descritta dai cluster trovati. Se uno dei 4 valori tende verso il 100%, più è probabile che quel cluster sia caratterizzato da un fattore per ora non considerato.

4. Risultati

Il modello di Deep Learning è stato allenato sul primo periodo di dati (4-11 Marzo 2020). Questi dati, sono stati divisi in set di training (72%), quello su cui è stato effettivamente allenato il modello, set di validation (18%), utilizzato per scegliere quando fermare l'allenamento della rete neurale, e set di test (10%), per valutare la performance del modello. Dopo aver allenato il modello, questo viene utilizzato per effettuare le predizioni per i tweets appartenenti ai periodi temporali seguenti, per poi calcolare le metriche da noi scelte.
Sul dataset di test, si raggiunge un'accuratezza del 93.7%. Questo ci suggerisce che il modello è in grado di replicare in maniera eccellente la tecnica di ETM per il primo periodo temporale.

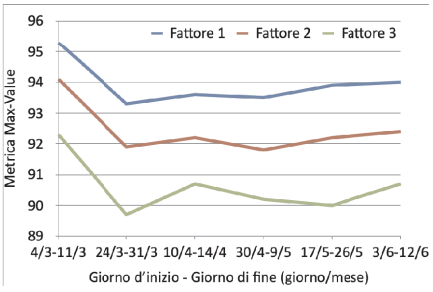


Figura 1. Valori metrica Max-Value per ogni fattore nei 6 periodi temporali considerati

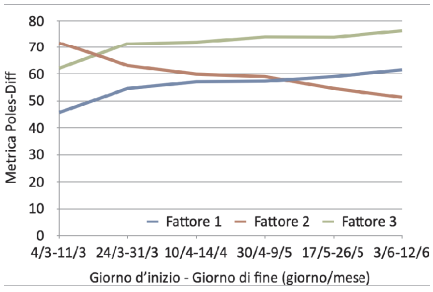


Figura 2. Valori metrica Poles-Diff per ogni fattore nei 6 periodi temporali considerati

La figura 1 mostra, per ogni fattore, il valore della prima metrica Max-Value. Possiamo constatare come i valori siano in ordine decrescente rispetto al numero del fattore; questo è da aspettarsi in quanto, in base alla tecnica di ETM, sappiamo il primo fattore è più importante del secondo, che è più importante del terzo, per spiegare la clusterizzazione. Notiamo inoltre una diminuzione netta nel secondo periodo, che potrebbe rappresentare la comparsa di un nuovo fattore, che non possiamo vedere se non riapplicando la tecnica di ETM.
La figura 2 mostra la metrica Poles-Diff calcolata per ogni fattore. Per il primo e il terzo fattore la metrica aumenta, indicando che i tweet risultano più polarizzati secondo essi. Per il secondo fattore vediamo invece come diventi meno importante per la categorizzazione; possiamo ipotizzare che un nuovo fattore stia comparendo e lo stia andando a sostituire.
La figura 3 mostra la frequenza percentuale per i 4 clusters più presenti nelle predizioni della rete. È interessante notare come, a partire dal secondo periodo, i primi 3 cluster mantengano all'incirca la stessa frequenza, decrescendo lentamente nel tempo, mentre il quarto cluster aumenti a loro discapito, andando a superare il 40%. Il quarto cluster rappresenta il sentimento riguardo il pericolo del Covid a livello nazionale. Ipotizziamo che l'aumento in frequenza sia collegato al fatto che dopo Marzo il pericolo è ormai nei confini nazionali.

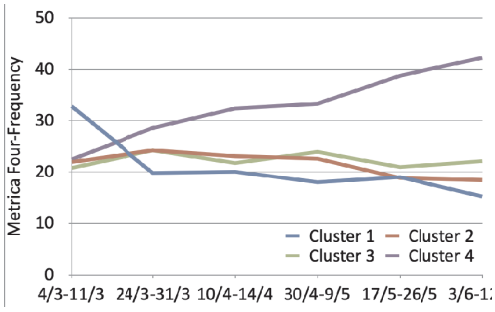


Fig 3. Valore metrics Four-Frequency per i 6 periodi temporali considerati

5. Conclusioni

Abbiamo mostrato come, con l'ausilio di un modello di Machine Learning, sia possibile imitare il funzionamento della tecnica di ETM. Grazie alle predizioni effettuate con tale modello e all'uso sapiente di metriche selezionate ad-hoc, siamo riusciti a valutare il cambiamento di opinione del pubblico in periodi temporali successivi a quello su cui è stata applicata la tecnica di ETM. Già nel secondo periodo, possiamo vedere un cambio repentino nell'espressione dei fattori, indicativo della capacità del primo e terzo fattore di rappresentare le dimensioni di senso presenti nella comunicazione e della decrescente rappresentatività del secondo fattore. Ciò suggerisce che a partire dalla fine di

marzo si venga a costituire un nuovo fattore peraltro confermato da analisi condotte in studi precedenti (Boccia Artieri et al., 2021a; 2021b). Infine, i risultati ottenuti suggeriscono come l'ETM potrebbe essere utilizzato come metodo per effettuare una prima classificazione automatica del testo per costituire il learning set. Infatti, la decodifica manuale del learning set può risultare onerosa in termini di tempi e costi nonché comportare problemi di classificazione in relazione alla grandezza dello stesso.

References

Boccia Artieri G., Greco F. and La Rocca, G. (2021). The flame of the Coronavirus in Italy. Looking for fear arousing appeal during the media hype period on Twitter. International Review of Sociology, 31(2), 287-309, https://doi.org/10.1080/03906701.2021.1947950

Boccia Artieri G., Greco F. and La Rocca, G. (2021). Lockdown and Breakdown in Italians' Reactions on Twitter during the First Phase of Covid-19. PACO, 14(1), 261-282. https://doi.org/10.1285/i20356609v14i1p261

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Giuliano L. and La Rocca G. (2010). *Analisi automatica e semi-automatica dei dati testuali: Strategie di ricerca e applicazioni* (vol. II). Led.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F. and Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management,* 51: 101934, https://doi.org/10.1016/j.ijinfomgt.2019.04.007.

Lebart L. and Salem A. (1994). *Statistique Textuelle.* Dunod.

Steinbach M., Karypis G. and Kumar V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 400, pp. 525–526. Boston.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Linguistic complexity and engagement in social media communication for cultural heritage

Massimo Guarino[1], Violetta Simonacci[2], Michele Gallo[1]

[1]University of Naples "L'Orientale" – mguarino@unior.it
[2]University of Naples Federico II – violetta.simonacci@unior.it

Abstract
This contribution has the dual goal of testing the usage of readability and lexical complexity tools in social media writing and analyzing the relationship between linguistic complexity, cultural topics, and social engagement. A corpus of tweets from the official Twitter page of the Italian Ministry of Culture has been stored together with available information on social engagement. Once the tweets have been properly processed with standard Natural Language Processing tasks, linguistic complexity is measured with eight readability and lexical diversity indicators developed outside the social media context. Topic modeling is then carried out through the LDA technique. An explorative study of cause-effect relationships between lexical complexity, topic grouping, and engagement measures is performed by using a PLS-PM mixed approach.

Keywords: corpus, NLP, readability indicators, tokenization, tweets, vocabulary

1. Introduction

The official Twitter page of the Italian Ministry of Culture (MiC) is a powerful resource of social media textual data on front-line cultural topics. A corpus, consisting of a large body of tweets from this page, and some additional measures such as likes and retweets were collected with the purpose of conducting a detailed study on readability, engagement, and cultural topics.
The analysis has been carried out in three working moments. The first phase includes a preprocessing stage and the implementation of a series of Natural Language Processing (NLP) tasks on all the documents of the corpus to clean, organize and lemmatize the data, preparing it for statistical analysis.
In the second working phase, topic and language complexity measures were computed. In detail, a topic analysis was conducted with the Latent Dirichlet Allocation (LDA) approach (Blei, 2012) to obtain indicators that differentiate documents based on dominant topic and common tokens incidence. Moreover,

a set of eight readability[1] indicators was calculated for each cleaned tweet. Some of these measures focus mainly on vocabulary choices, while others are devised to give a better grasp of language richness[2]. Testing the readability of texts in the social media context can help further explain how the use of the language on mass media communication platforms can impact organizations' success with respect to their social engagement policies.

In the final working phase of this project, the relationships among the considered measures, both directly observed and computed, are analyzed via PLS-PM. In detail, the following questions were considered when building the model: 1) What is the relationship between the different computed measures, and can they be synthesized in meaningful compound constructs? 2) Are the readability indicators a good measure of language structure and richness in a social media setting? 3) What role do language complexity and topic indicators play in social engagement with respect to cultural communication?

It is important to note that the PLS-PM model is used for theory building rather than for confirming hypotheses. In this sense, results are viewed from an exploratory standpoint as a first step in the search for possible theories and connections of interest without assumptions (Henseler, 2008). For this reason, the well-known limits of the procedure were not addressed in detail here (see Rönkkö et al., 2016 for an in-depth discussion on PLS shortcomings).

The constructed dataset is quite rich and complex. Details on preprocessing and indicators construction are provided in Section 2. A quick overview of the PLS-PM modeling tool is presented in Section 3. Preliminary results of the model are presented in Section 4.

2. Data and readability indicators

We built the textual dataset based on the tweets of the MiC collected on its Twitter official page (about 2,300 tweets). For each tweet, we also collected the number of likes and retweets and its creation date. Furthermore, we performed a series of NLP tasks in order to prepare all the documents for the computation of indicators.

Preprocessing steps were necessary to remove textual noises and extra whitespaces to perform better tokenization (i.e., splitting each text into a set of sentences and sentences into a set of words, the "word-tokens"). Once preprocessed, word-tokens were lowercased to put all the texts on a level playing

[1] Readability can be defined as "what makes some texts easier to read than others" (DuBay, 2004).

[2] Vocabulary knowledge, including lexical diversity and richness, are principal aspets in reading comprehension (Heilman et al., 2008; Gooding et al., 2021). Moreover, the importance of lexical richness has been investigated also in readability systems (Vajjala and Meurers, 2012; Xia, 2019)

field and two other relevant NLP tasks were performed using Stanza[3]: Part of Speech (PoS) tagging and lemmatization.
Pre-processed tweets were clustered into 17 groups via LDA, an unsupervised machine learning technique providing highly interpretable topics based on Bayesian algorithms and Dirichlet probability distributions (Blei, 2012). Given the nature and length of the tweets, we chose to focus only on some content words (adjectives, nouns, and verbs) for topic detection. The number of groups was chosen by maximizing topic coherence, a reliable measure used to determine the readability/interpretability of topics (Tijare and Rani, 2020). Two topic indexes were built: Top1, which groups documents based on the incidence of common tokens (a higher value is associated with a smaller percentage of such tokens), and Top2 which expresses the percentage of dominant topic.
Successively, eight readability/lexical richness indicators were computed for each tweet. To start, we define: TTR[4], the type/token ratio, namely the ratio between the number of different words in a text and the total amount of its word-tokens; HAP, the percentage of hapaxes (words which occur only once in a text); and CONT the percentage of content words[5].
A fourth fundamental indicator is given by the *Gulpease* index (Gulp), which is of particular interest because it was developed in reference to the Italian language and educational system (Lucisano and Piemontese, 1988). It can be described by the following formulation:

$$Gulpease = 89 + \frac{(300 \times sentences - 10 \times letters)}{number\ of\ words} \quad (1)$$

The index ranges between 0 (un-intelligible) and 100 (maximum readability) and it is inversely correlated with the degree of educational level. Documents with a

[3] While for sentence recognition and tokenization we used Stanza, the NLP tool developed within the Stanford NLP research group (Manning et al., 2014; Peng et al., 2020), for words and letters tokenization we used regular expressions, which are sequences of characters that define a search pattern (Aho, 1991). Word tokenization refers only to words, thus excluding symbols and punctuations from the computation.
[4] Given that tweets do not greatly differ in terms of texts length, the ratio between types and tokens can be considered stable.
[5] There are 3 main class words in the universal part of speech tagging framework (i.e., the standard part of speech tagset for many languages) (Murphy 2010, Petrov et al., 2011): open class words (the most important class comprising the so called content words, that is to say words giving meaning to a text as adjectives, nouns, verbs (other than auxiliary verbs), proper nouns, adverbs and interjections), closed class words (this is the set comprising the so called function words, that is to say prepositions, auxiliary verbs, conjunctions, pronoun, etc..), and other words (everything not comprised within the previous classes such as punctuation, symbols and other misclassified elements).

low level of readability (i.e. with an index lower than 60) are hardly understood by native speakers with junior high school education; a value lower than 40 means that documents are not easily understandable by those with high school education.

The last four indicators are connected to the use of the Italian *Base* Vocabulary (De Mauro, 1980; De Mauro and Chiari, 2016), namely a collection of the first high-frequency words (about 7,000 lemmas) of the written Italian language which are greatly understandable also by those who are less educated. The very first 2,000 lemmas of this collection constitute the Italian *Fundamental* Vocabulary. In this perspective, we computed the share of content words present in the base vocabulary (Dic0) and fundamental vocabulary (Dic1). To avoid biases due to the tokenization of the tweets, we also computed two more indicators with respect to the complete text, and not just the content words (labelled respectively Dic2 for the base vocabulary, and Dic3 for the fundamental one).

In addition to the described indicators, two simple volume measures were also computed to give general size information: the total number of words (Vol1) and the total number of words without repetitions (Vol2). Lastly, the Engagement measures considered in the model were extracted directly from the tweets and include the number of likes (Hot1) and the number of retweets (Hot2)

3. PLS-PM method

To investigate relationships among the different types of measures discussed, a multivariate procedure capable of unveiling latent constructs and probing linear structures can be very helpful. An exploratory PLS-PM approach appears to be a suitable choice as no specific assumptions were made. The procedure identifies latent constructs for the manifest variables, represented in our data by readability indicators, volume, topic, and engagement measures, while simultaneously verifying the relationships between the constructs. PLS-PM can be seen as a parallel combination within an iterative scheme of path analysis to estimate the relationship among latent variables (structural or inner part of the model) and a factor analysis to obtain these latent measures from the observed variables (measurement or outer part of the model).

The structural or inner model can be described in brief by the following formulation where the path coefficient β_{mi} expresses the causal effect between the generic endogenous latent variables ξ_j and the $[1, \dots, m, \dots, M]$ exogenous latent variables ξ_m:

$$\xi_j = \sum_{m=1}^{M} \beta_{mj}\, \xi_m + \tau_j \tag{2}$$

The value τ_j denotes the structural error. The measurement or outer model for linking the manifest variables to the latent constructs can be either reflective, formative, or mixed, on the basis of theoretical or empirical reasons. In the reflective blocks, the manifest variables are intended as the manifestations of the latent construct. Let us consider the q-th block referring to the $[1, \dots, p, \dots, P]$ manifest variables x_{p_q}. The relationship between the manifest variables and their corresponding latent construct ξ_q can be formulated as:

$$x_{pq} = \lambda_{pq}\xi_{pq} + \varepsilon_q \tag{3}$$

Here $\lambda_{pa}$is the simple regression coefficient between the manifest variable and the latent variable which is denoted as loading.

In formative blocks, the latent variables are seen as the best predictors of the manifest variables under the constraint of explaining the relationships between latent adjacent measures. In this case, the latent construct is expressed as the linear combination of the manifest variables with associated weights w_{p_q}in the following manner:

$$\xi_q = \sum_{p=1}^{p_q} w_{p_q} x_{p_q} + \delta_q \tag{4}$$

Here a mixed scheme is used for the measurement model. Parameter estimation is carried out with an iterative process that allows the simultaneous estimation of latent variables, outer and inner weights. After initializing of outer weights, the procedure conducts alternating inner and outer estimating steps through a system of linear equations by differentiating outer weights computation for reflective and formative blocks, until convergence is reached. For a detailed illustration of the PLS-PM procedure please refer to Tenenhaus et al., 2005.

4. Preliminary findings and discussion

The model input includes 14 manifest variables arranged in $q=6$ blocks, consisting of 8 language complexity indicators, 2 volume, 2 topic, and 2 engagement measures. Three blocks have reflective relations and include: 1) Volume {Vo1; Vol2}; 2) Dictionary {CONT, Dict0, Dict1, Dict2, Dict3}; and 3) Languange Richness {TTR, HAP}. Here, the reliability of the constructs is demonstrated by the high Cronbach's alpha values, equal to 0.976, 0.895, and 0.961 respectively. All loadings are acceptable except for Dict3, which records a value of 0.197. The remaining blocks have formative ties and include: 4) Topic {Top1, Top2}; 5) Structure {Gulp}; and 6) Engagement {Hot1; Hot2}.

The structural model estimates 1) the relationships between the exogenous latent variables Volume, Dictionary, and Richness with Structure and Topic and 2) the relations between the variables Structure and Topic with Engagement. To assess results, beta coefficients are reported in Table 1:

Table 1. Structural model Path Coefficients

	Structure	Topic	Engagement
Volume	0.008	-0.117	-
Dictionary	0.163***	0.597	-
Richness	-0.073***	0.619	-
Structure	-	-	0.344
Topic	-	-	-0.49

R^2 determination coefficients are quite low with the exception of Topic, where it corresponds to 0.774. This is, however, expected as many other aspects impact engagement on social media which are not the focus of this analysis. In addition, very few path coefficients are statistically significant. This is also not surprising nor problematic at this stage as the model was carried out with a heuristic approach to explore rather than to infer.

$$\xi_q = \sum_{p=1}^{p_q} w_{p_q} x_{p_q} + \delta_q \tag{3}$$

These preliminary results, which need additional evidence by extending the corpus and refining indicators, show how an Engagement increase could be brought by a higher value of Structure, which stems from the *Gulpease* readability index. On the other hand, as the topic scores increase (the topic becomes more specific), Engagement may be reduced. A positive connection also links Dictionary with Structure and with Topic, and Richness with Topic.

References

Aho A. V. (1991). Algorithms for finding patterns in strings. *Handbook of theoretical computer science (vol. A): algorithms and complexity*. MIT Press, Cambridge, MA.

Blei D. M (2012). Probabilistic topic models. *Communications of the ACM,* 55(4): 77–84

Manning C. D., Surdeanu M., Bauer J., Finkel J. R., Bethard S. and McClosky D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*.

De Mauro T. and Chiari I. (2016). Il Nuovo vocabolario di base della lingua italiana. https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

De Mauro T. (1980). *Guida all'uso delle parole: Come parlare e scrivere semplice e preciso: Uno stile italiano per capire e farsi capire*. Editori Riuniti.

DuBay W. H. (2004). *The Principles of Readability*. *Costa Mesa: Impact Information*, vol. (76).

Gooding S., Berzak Y., Mak T. and Sharifi, M. (2021). Predicting Text Readability from Scrolling Interactions. *arXiv preprint arXiv:2105.06354.*

Henseler J. (2018). Partial least squares path modeling: Quo vadis? *Quality & Quantity*, *52*(1) : 1–8.
Heilman M., Collins-Thompson K. and Eskenazi M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications.* pp. 71–79.
Lucisano P. and Piemontese M. E. (1988). GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31): 110–124.
Murphy M. L. (2010). *Lexical meaning*. Cambridge University Press.
Peng Q., Zhang Y., Zhang Y., Bolton J. and Manning C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
Petrov S., Das D. and McDonal R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086.*
Rönkkö M., McIntosh C. N., Antonakis J. and Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management,* vol(47) : 9–27.
Tenenhaus M., Esposito V., Chatelin Y. M. and Lauro C. (2005). PLS path modeling. *Comput Stat Data Anal,* vol(48): 159–205.
Tijare P. and Rani P. J. (2020). Exploring popular topic models. *Journal of Physics: Conference Series,* 1706(1): 012171. IOP Publishing.
Xia M. 2019. *Text readability and summarisation for non-native reading comprehension*. Ph.D. thesis, University of Cambridge.

Sales forecast and electronic word of mouth: the power of feelings

Domenica Fioredistella Iezzi, Roberto Monte

University of Tor Vergata – stella.iezzi@uniroma2.it
roberto.monte@uniroma2.eu

Abstract
Research has shown that word of mouth (WOM) among consumers is more effective in traditional decision-making than any marketing tool, such as personal sales or advertising media. With the development of the Internet, electronic WOM has replaced traditional word of mouth, becoming an informal and effective exchange of person-to-person communication between a perceived non-commercial communicator and a contribution regarding a brand, product, organization, or service company.
The paper aims to improve the predicted sales using the sentiment component from social media posts and reviews. To this end, we have introduced a suitable e-WOM state-space model, inspired by the autoregressive sentiment aware (ARSA) models. Hence, we have tested our SSM on two different products: the sales of Gentilini's Osvego biscuits and the sales of electric cars in the US market. The sentiment has been measured using two corpora of Tweets, one in Italian for Gentilini biscuits and one in English for the electric car market.

Keywords: State Space Models, ARSA models, sales forecast, sentiment analysis.

1. Introduction

In recent years, people frequently post their opinion of purchased products on social platforms, such as Google and Twitter. These user-generated posts, reviews, tweets tell other users about the quality of the products reviewed. According to Brown and Reingen (1987) and Katz and Lazarfeld (1955), word of mouth (WOM) is more effective in consumer decision-making than other traditional marketing tools, such as personal sales or advertising media. Recently, the Internet has played a significant role in exchanging information, and electronic WOM (e-WOM) has become the primary informal exchange of information between people. e-WOM has become an informal exchange of person-to-person communication between a perceived non-commercial communicator and a recipient regarding a brand, product, organization, or

company service (Bartschat *et al*., 2022; Fronzetti *et al*., 2021). Wright (2009) points out that "online customer opinions have become a virtual currency that can make or break their products". Duan *et al*. (2008) claim that "WOM has been recognized as one of the most influential resources of information transmission since the beginning of society, especially for experience goods"- Several studies have confirmed that e-WOM significantly impacts product sales. For example, Chen et al. (2011), using the reviews of the online seller Amazon.com., designed three longitudinal, quasi-experimental field studies to examine three issues regarding the two types of interaction: 1. their differential impact on product sales, 2. their lifetime effect, and 3. their interaction effects. An interesting result is that negative e-WOM is more influential than positive.

Dual and Whinston (2008) examine the persuasive impact of the awareness effect of online user reviews on movies' daily box office performance. They developed a simultaneous equation system to explore the relationship between online reviews and the box office revenues of movies. Hand and Judgeb (2012) use monthly data for the United Kingdom, 1 January 2004 to 31 December 2008, to examine various forecasting models incorporating Google Trends search information. They have found evidence that Google Trends data on searches relevant to cinema visits, as opposed to searches for specific films, do have the potential to increase the accuracy of cinema admissions forecasting models. Park and Chung (2012) discussed how the number of times a comment is shared on Twitter is related to sales volume. The paper aims to understand if and how consumer sentiment affects sales dynamics. We use aggregate sales time series because they contain trend and seasonal patterns. They provide a good testing ground for comparing forecasting methods; because companies can benefit from more accurate forecasts (Ramos *et al*., 2015). We have introduced a new idea to build an SSM to forecast sales using sentiments analysis from social media posts or reviews. We have used the monthly time series for experimentation with the models: on the sales of electric cars in the USA market and the sales of the Osvego biscuits of the Italian Gentilini Company. Sentiment analyses have been conducted on two corpora: one in English (data on electric cars) and the other in Italian for biscuits, trying to standardize the analysis strategies as much as possible. The rest of the paper is organized as follows. Section 2 presents data and our e-WOM state-space model. In Section 3, we discuss the data and the first results.

2. The e-WOM state-space model

Given a monthly time series of volume sales data for a product, $y \equiv (y_k)^t{}_{k=0}$, and a collection of monthly tweets pertaining the product, $l \equiv (l_k)^t{}_{k=0}$, we attempt to construct a simple sales forecasting model that considers the influence of the tweets for predicting the future sales of the product. This consists of a state-space model in which the time series *y* and *l* are interpreted as the result of noisy observation of an underlying unobservable stochastic process. Inspired by the so-called ARSA model (Liu *et al.,* 2009; Yu *et al.*, 2012) and Hyndman's contributions to the SSM theory (Hyndman *et al*., 2008), we build an SSM in which sales are driven by a simple linear process around trend, seasonality, and sentiment levels.

Our e-WOM state-space model has the following form:

$$X_t = FX_{t-1} + U_t \tag{1}$$

$$Y_t = GX_{t-1} + V_t \tag{2}$$

where $\mathbf{X}_t \equiv (\boldsymbol{S}_t, b_t, l_t) \in \mathrm{R}^{2+m}$ is a partially unobservable state vector with components $\boldsymbol{S} \equiv (\boldsymbol{S}_k)^t{}_{k=0} \in \mathrm{R}^m$ [resp. $(b_k)^t{}_{k=0} \in \mathrm{R}$, resp. $l \equiv (l_k)^t{}_{k=0} \in \mathrm{R}$], which represents the seasonality component of *m* month-period [resp. trend component, resp. sentiment level component]; $\mathbf{Y}_t \equiv (y_t) \in \mathrm{R}^2$ is an observation vector with components the above-mentioned sentiment level and the sales volume; **F** and **G** are the transition and design matrices of the model, respectively, to represents the linear relationships among the components of **X** and **Y**; in the end, **U** and **V** are suitable jointly Gaussian distributed noise matrices, to represent the noises affecting the components of **X** and **Y**. Differently than a typical ARSA model, in our approach the trend and seasonality are estimated using the Kalman filter. Unlike Hyndman's approach, the sentiment appears in both the state and observation components.

3. Data and results

We have used two monthly sales time series to test our e-WOM state-space model: the Osvego Biscuits sales, from January 1, 2013, to December 31, 2020, and the USA electric cars sales, from January 1, 2012, to December 31, 2019. We have collected simultaneously two chronological corpora, for the same periods' series of sales, on Twitter for Osvego biscuits (no. 998) and electric cars in the USA market (no. 100000).

We propose two different methods to estimate the sentiment: lexical and machine learning approaches. In the linguistic approach (Birjali et al., 2021), we have built a vocabulary in R based on the Data Bank Zampolli (Italian Sentiment Lexicon), created by the National Research Center of Pisa. It is freely downloadable from the Internet. The Italian Sentiment lexicon was developed semi-automated from ItalWordNet, starting from a list of 1.000 manually checked seeds. It contains

24,293 lexical entries, annotated with positive/negative/neutral polarity. We have also carried out the stemming of corpus for the Italian tweets and the usual cleaning operations. In the Machine Learning approach, we have used the VADER package (Valence Aware Dictionary for Sentiment Reasoning), which allows us to measure both the polarity (positive/negative) and the intensity (strength) of emotions (Hutto, Gilbert, 2015). VADER sentiment analysis maps lexical characteristics to emotional powers known as sentiment scores. Moreover, it emphasizes capitalization and punctuation and makes it particularly performant on texts from social media such as Twitter. Seasonality or special events such as holidays and promotions often affect sales. However, lately, it is possible to observe that, once seasonality and particular events are eliminated, consumer products are still affected by the trend in sales which seem to be linked to public sentiment on the product. Our e-WOM state-space model incorporates the sentiment level, improving sales forecasts.

$$X_t = FX_{t-1} + U_t \tag{1}$$

$$Y_t = GX_{t-1} + V_t \tag{2}$$

where $\mathbf{X}_t \equiv (\boldsymbol{S}_t, b_t, l_t) \in \mathrm{R}^{2+m}$ is a partially unobservable state vector with components $\boldsymbol{S} \equiv (\boldsymbol{S}_k)^t{}_{k=0} \in \mathrm{R}^m$ [resp. $(b_k)^t{}_{k=0} \in \mathrm{R}$, resp. $l \equiv (l_k)^t{}_{k=0} \in \mathrm{R}$], which represents the seasonality component of *m* month-period [resp. trend component, resp. sentiment level component]; $\mathbf{Y}_t \equiv (y_t) \in \mathrm{R}^2$ is an observation vector with components the above-mentioned sentiment level and the sales volume; **F** and **G** are the transition and design matrices of the model, respectively, to represents the linear relationships among the components of **X** and **Y**; in the end, **U** and **V** are suitable jointly Gaussian distributed noise matrices, to represent the noises affecting the components of **X** and **Y**. Differently than a typical ARSA model, in our approach the trend and seasonality are estimated using the Kalman filter. Unlike Hyndman's approach, the sentiment appears in both the state and observation components.

Table 1. Model Statistics – Dataset OSVEGO

Models	log Lik	AIC	AICc	BIC	MSE	LB test	SW test	DP test
ETS - AAA (forecast)	-30.73	95.46	105.33	135.96	0.0270	0.766	0.279	0.799
SP (MARSS)	-5.12	40.23	47.73	10.23	0.0638	2.2e-16	0.729	0.854
e-WOM SP Comp (L. appr.)	-22.76	81.51	86.36	45.51	0.1039 (0.0346)	(0.569)	(0.497)	(0.371)
e-WOM SP Neg. (L. appr.)	5.83	24.33	29.18	-11.67	0.0577 (0.0411)	(1.82e-13)	(0.804)	(0.833)
e-WOM SP Comp. (ML appr.)	48.09	-60.18	-55.32	-96.18	0.0318 (0.0311)	(0.693)	(0.219)	(0.116)
e-WOM SP Neg. (ML appr.)	185.74	-335.47	-330.62	-371.47	0.0183 (0.0356)	(0.914)	(0.793)	(0.773)

*The results in brackets concerning e-WOM SP models are referred to as the volume sales component of the OSVEGO Data, which corresponds to the observed part in ETS and SP models. LB Test= Lyuag-Box Test, SW= Shapiro Wilks normality Test, DP Test=D'Agostino Pearson Test.

Table 1 shows the main model statistics of dataset OSVEGO calculated on six different models. In the first two models, ETS (Error Time Seasonality – Hyndman et al., 2008) and SP (State Space model), we only use sales data to make predictions, while in the other four models, we add the sentiment component. In e-WOM SP (State Space Model), we estimate sentiment with a double approach: lexical and Machine Learning (ML), using compound and negative sentiment. Note that negative sentiment improves the model more than the compound. This is a somehow surprising result. Because of this, in forthcoming developments of the models, we are planning to introduce separate positive and negative components of the sentiment better to analyse the role of emotion in consumer choices. In our results, the best performance comes from the e-WOM SP negative ML approach among the models. We developed Time Series Analysis and State Space Models with R statistical software (forecast and MARSS libraries).

References

Bartschat M., Cziehso G. and Hennig-Thurau T. (2022). Searching for word of mouth in the digital age: Determinants of consumers uses of face-to-face information, internet opinion sites, and social media, *Journal of Business Research*, 393-409

Birjali M, Kasri M and Beni-Hssane A (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends, *Knowledge-Based Systems*, 226 (2021), 107134, https://doi.org/10.1016/j.knosys.2021.107134.

Brown J. J. and Reingen P. H. (1987). Social Ties and Word-of-Mouth Referral Behavior. In *Journal of Consumer Research*, 14(3), 350

Chen Y, Wang Q. and Xie J (2011). Online social interactions: a natural experiment on word of mouth versus observational learning. *J Mark Res* 48(2):238–254

Duan W., Gu B. and Whinston. A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. *Decision Support Systems* 45 (4): 1007-16.

Durbin J. and Koopman S. J. (2012). *Time Series Analysis by State Space Methods*, Oxford University Press (2012).

Fronzetti Colladon A., Gloor P. and Iezzi, D. F. (2021). Editorial introduction: The power of words and networks. *International Journal of Information Management*, 51.

Hand C. and Judgeb G. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. *Applied Economics Letters*, 19: 1051–1055.

Hutto C.J. and Gilbert E. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM 2014.

Hyndman R.J., Koehler A.B., Ord J.K. and Snyder R.D. (2008). *Forecasting with exponential smoothing: the state space approach*, Springer.

Katz E., Lazarsfeld P. F. (1955). Personal Influence, *The Free Press*, Glencoe

Liu Y., Yu X., Huang X. and An A. (2009). *Blog Data Mining: The Predictive Power of Sentiments*. 10.1007/978-0-387-79420-4_13.

Messaoudi C., Guessoum, Z. and Ben Romdhane, (2022). Opinion mining in online social media: a survey. *Soc. Netw. Anal. Min.* **12,** 25.

Park J. Y. and Chung C.W. (2012). When daily deal services meet Twitter: understanding Twitter as a daily deal marketing platform. *WebSci '12: Proceedings of the 4th Annual ACM Web Science Conference,* June 2012 Pages 233-242
Ramos P., Santos N. and Rebelo R.D. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-integrated Manufact*uring, 34: 151-163.
Wright A (2009). Our sentiments, exactly. *Commun ACM* 52(4):14–15
Yu X., Liu Y., Huang, X. and An A. (2012). Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain. *IEEE Trans. Knowl. Data Eng*.. 24. 720-734. 10.1109/TKDE.2010.269.

Cree Legends and Narratives au prisme d'IRaMuTeQ

Julie-Pier Jean, Sylvia Kasparian

Université de Moncton – julie-pierjean@hotmail.ca
sylvia.kasparian@umoncton.ca

Abstract
What are the lexical worlds and myths conveyed in Cree stories and legends? This is the question we try to answer by analyzing the tales in the volume *Cree Legends and Narratives: From the West Coast of James Bay*, by Ellis C. Douglas and Roman Yereniuk. Using the divisive hierarchical clustering (DHC) proposed by Reinert (1993) and implemented in the IRaMuTeQ software (Ratinaud 2009), we attempt to apply this method and analyze transcribed oral discourse in order to validate the contribution of this method of analysis and to illuminate our knowledge of the values and the conceptions of the world of the Cree peoples.

Keywords: Reinert method, hierarchical top-down classification, universe of discourse, IRaMuTeQ, Cree stories, myths and symbols, Aboriginal peoples of Canada

Résumé
Quels sont les univers du discours et les mythes véhiculés dans les contes et légendes Cris ? C'est la question à laquelle nous essayons de répondre en analysant les contes du volume Cree *Legends and Narratives: From the West Coast of James Bay,* de Ellis C. Douglas et Roman Yereniuk à l'aide de la classification hiérarchique descendante proposée par Reinert (1993) tel qu'implémenté dans le logiciel IRaMuTeQ (Ratinaud 2009). C'est donc une application de cette méthode à un type particulier de discours oral transcrit, afin d'analyser et valider l'apport de cette méthode d'analyse à nos connaissances sur les valeurs et la conception du monde des Cris.

Mots clés: méthode Reinert, classification hiérarchique descendante, univers de discours, IRaMuTeQ, contes cris, mythes et symboles, Autochtones du Canada

1. Introduction

Les Cris, composent le « peuple autochtone le plus peuplé et le plus largement réparti au Canada » et occupent un territoire qui va de l'Alberta au Québec en passant par les plaines de l'Alberta et de la Saskatchewan (Preston 2012). Ils se

distinguent des autres peuples entre autres par leurs histoires et leurs cultures uniques, riches en enseignements. De traditions orales, ces enseignements étaient jadis transmis de génération en génération par le biais des contes et légendes qui véhiculent des symboles, mais aussi plusieurs mythes fondateurs qui perdurent. Ces mythes et ces symboles culturels forgent l'imaginaire culturel des Cris et sont le reflet de leurs perceptions du monde. Ce sont essentiellement ces représentations sociales culturelles qui nous ont intéressés et qui nous ont poussé à étudier les contes cris afin d'examiner comment les contes et légendes cris sont des moyens privilégiés de transmission et de partage de connaissances profondément ancrées dans la culture.

Selon Rocher (1968 : 117) la culture est un « univers mental, moral et symbolique, commun à une pluralité de personnes, grâce auquel et à travers lequel ces personnes peuvent communiquer entre elles, se reconnaissent des liens… » et qui leur permet de se sentir individuellement et collectivement membres d'une même entité qui les dépasse et qu'on appelle un groupe ou une société. Ainsi, les mythes véhiculés dans une société forment un ensemble et « tissent tout autour du monde une trame logique en perpétuelle évolution » (Descola, 2008 : 5). Ainsi, les mythes participent à la construction d'une vision du monde, de représentations qui sont une « forme de connaissance socialement élaborée et partagée, ayant une visée pratique et concourante à la construction d'une réalité commune à un ensemble social » (Jodelet 1997 : 53). Dès lors, l'analyse des mondes lexicaux communs à un groupe, véhiculés par les contes fondateurs d'une culture peut nous donner des indices des représentations sociales que se fait le groupe d'un objet du monde social. En effet, selon Max Reinert, dans un corpus collectif, un monde lexical serait indicateur d'un espace de référence commun à un groupe et « l'indice d'une forme de cohérence liée à l'activité spécifique du sujet-énonciateur » (Reinert, 1993; 13).

La méthode de la classification hiérarchique descendante (Reinert 1990) propose une représentation de ces mondes lexicaux sous la forme de tableaux de classification obtenus par croisement des unités de contexte (ou segments) et des lexèmes d'un corpus. L'hypothèse à la base de cette méthode est que « dans la mesure où une représentation collective exprime une certaine régularité de structure dans une classe de représentations singulières […] cette régularité est due aux contraintes de ce que nous appelons "un monde" » (Reinert, 1993 : 29-30). La prise en compte de la fréquence et de l'environnement des formes d'un corpus permet non seulement de relever les formes lexicales les plus propices à constituer des indices de représentations sociales, mais aussi de définir ces formes lexicales en fonction de leur cotexte. Nous souhaitons donc à partir de ces analyses identifier les éléments représentationnels (valeurs, mythes, symboles) présents dans les contes Cris.

2. Corpus et Analyse

Afin d'atteindre notre objectif, nous avons choisi d'analyser par la classification hiérarchique descendante de type Reinert (1990) implémentée dans Iramuteq (Ratinaud, 2009), une série de 63 contes cris tirés de *Cree Legends and Narratives : From the West Coast of James Bay* des auteurs Ellis C. Douglas et

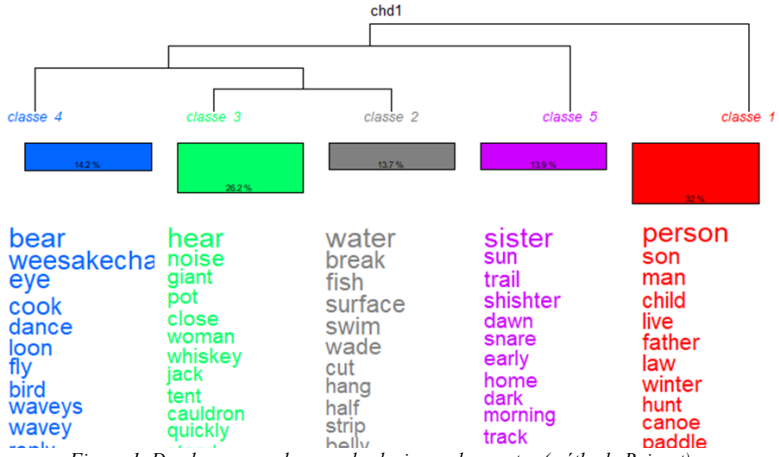


Figure 1. Dendogramme des mondes lexicaux des contes (méthode Reinert)

Roman Yereniuk (2014) transcrits et traduits en anglais, à partir d'enregistrements sur cassette de la narration de seize aînés cris.

La longueur des contes varie entre 50 et 400 mots, scindés en 1188 segments pour un total de 43 958 occurrences (2791 formes différentes) regroupées en 2142 lemmes, soit 1747 formes actives (mots pleins) et 395 formes supplémentaires (mots outils). Le résultat de cette analyse est représenté par la figure 1 (ci-dessus) qui organise ces récits en 5 classes ou mondes lexicaux.

2.1. La classe 1 : Représentations du rôle de l'homme cri

Dans la classe 1 (en rouge, 32% des segments), on aborde la dynamique familiale avec les formes « son », « child », « father », « law » (pour « brother in law » ou « son in law »). On remarque aussi, en plus des formes relevant des relations familiales entre les hommes du clan, un champ lexical qui semble indiquer la présence d'un univers relié aux activités entre père et fils telles que la chasse et le canoë principalement l'hiver et l'été représentés par les formes « winter », « hunt » « canoe », « paddle » « summer » qui, à se fier à cette liste de formes en contexte, jouerait un rôle d'avant-plan dans les représentations de ces relations familiales. Cette classe, en plus de montrer l'importance des liens familiaux, fait référence aux pratiques cries qui sont transmises de pères en fils puisque la chasse et l'utilisation de canoë pour se déplacer sont principalement réservées aux hommes.

D'après Ohmagari et Berkes (1997) dans les communautés cries recensées, tous les hommes suffisamment robustes prenaient part à la chasse à l'orignal, au caribou, à l'oie, au castor. Le mot « winter » est cooccurrant à « hunt », car les Cris pratiquent surtout la chasse à l'orignal durant l'hiver. Les chasseurs cris privilégient cette saison, car le ventre des orignaux traine dans la neige et qu'ils peuvent moins bien se déplacer (Feit 2000). Les chasseurs cris attendent donc les bonnes conditions météorologiques pour procéder à la chasse à l'orignal qu'ils pratiquent uniquement durant les "journées à orignaux" ce qui leur permet de les tuer sans les faire souffrir. Par ailleurs, leur organisation de la chasse à l'oie est particulièrement intéressante. En effet, les Cris suivent la migration des oies, ce qui requiert selon Sayles et Mulrennan (2010) « an elaborate system of social coordination that includes rotational strategies of hunting to ensure optimal harvests while reducing goose population disturbance ». Élément crucial pour les Cris, cette pratique fait partie des traditions ancestrales. Au printemps, quand la neige fondait et que le déplacement n'était pas possible, la chasse à l'oie leur permettait de subsister (Grand Council of the Crees, 2019). Encore aujourd'hui, la chasse permet aux familles et à la communauté crie de se réunir ; c'est aussi une opportunité pour les jeunes chasseurs d'apprendre des aînés comment construire des caches efficaces, attirer les oies à l'aide d'appels et d'appelants, préparer les oies pour les cuire sur un feu à ciel ouvert. Ces compétences sont profondément enracinées et transmettent un sens du travail d'équipe et le respect de la famille. (Grand Council of the Crees, 2019)

2.2. La classe 2 : Univers de l'eau et de la pêche

Quant à la classe 2 (en gris, 13,7% des segments), elle couvre l'univers de l'eau et de la pêche avec des mots tels que « water », « break » et « surface » dans le sens de briser la surface de l'eau glacée, ou « swim » et « cut », dans le sens de couper la glace ou couper un poisson. Voici un segment caractéristique de cette classe: « sight of it, the **fish** were in fact already **breaking** the **surface** in ripples again at that, he jumped straightway into the **water**, he began to **swim** ». Comme le dit Royer (2015) en plus d'être un moyen de subsistance des communautés cries, la pêche est considérée comme une partie intégrante de l'identité du Cri. D'après Berkes (1979), les Cris ne font pas de pêche commerciale, ils pratiquent strictement la pêche domestique où les prises sont réparties entre les membres de la communauté. Le mot « net » renvoie notamment à l'utilisation principalement des filets maillant (en anglais « gillnets ») pour pêcher. Tout comme pour la chasse, les Cris suivent la migration des poissons et les saisons pour savoir quand aller pêcher (Berkes, 1979), car pour eux il y a trois principaux temps de pêche : « August First Rapids fishery », « pre-freezup fishery » et « ice-fishery ». Les mots « break » et « cut » renvoient aux gestes techniques pour percer la glace, mais associés à « strips » ce dernier s'inscrit dans la fabrication

des filets. Le mot « belly » quant à lui fait référence dans les contes au ventre du « giant fish » qui a été coupé par la sœur de Chahkabesh et fait allusion à l'apprêtage du poisson exercé principalement par les femmes cries qui se chargent de vider les entrailles du poisson par une coupe ventrale. Finalement, les mots « hang » et « twisted » font référence à une méthode de cuisson du poisson par les Cris qui le pendent et le maintiennent sur des *bonask ou* bâtons coupés « cut » en pointe, dont on a enlevé l'écorce (Aubry et Lievre, 2016).

2.3. La classe 3 : Représentations du rôle de la femme crie

La classe 3 (en vert, 26,2% des segments) représente l'univers du travail domestique de la femme crie dans le campement. On y retrouve des mots associés à la préparation de la nourriture avec les mots « pot », « tent », « cauldron » (chaudron) et « fire ». Un segment caractéristique de cet univers est « However that **pot** stopped boiling the fat spread completely over the top but that **giant women** heard her big **pot** stop boiling suddenly. She glanced over, she looked at the **cauldron** ». Le rôle principal de la femme sur le campement, assurer toutes les étapes de la préparation de la nourriture (Davis, 2017), est complémentaire à celui de l'homme qui a la responsabilité d'aller chasser et d'assurer la nourriture pour le clan. Néanmoins, il arrive souvent, d'après Davis, que la femme crie parte aussi à la chasse et si ses enfants n'ont pas de père, elle est aussi en mesure de leur apprendre les rudiments de la chasse. Dans cette classe on retrouve aussi l'adjectif « giant » qui est lié au mot « women ». L'un des contes parle de deux femmes géantes qui grattent des peaux de castors et qui ramassent du bois dans la forêt, activités que pratiquaient les femmes cries. On rencontre également les mots « noise » et « hear », qui renvoient aux gros bruits que font les femmes géantes lors de la réalisation de ces tâches. On retrouve un autre personnage mythique dans cette classe, il s'agit du héros Whisky Jack ou Chahkabesh, un farceur qui joue des tours. Selon Preston (2012) les Cris croient qu'une personne peut apprendre des leçons de vie importantes de ces filous. On a également le mot « fire » associé à « women » qui s'explique par le fait que les femmes sont représentées comme les « fire keepers ». Selon Davis (2017 : 21) dans la culture crie la femme est associée à la chaleur et le confort que procure le feu au centre du foyer. En effet "Iskwew" est le terme utilisé pour femme et dérive du mot feu "iskwuptew", celui qu'on entretient au centre du tipi/wigwam. En somme, la femme crie possède un rôle central dans l'organisation sociale de la communauté.

2.4. La classe 4 : Univers des relations avec la faune

La classe 4 (en bleue,14,2% des segments) fait principalement référence aux animaux avec les mots « bear », « loon » (huard), « fly », « bird » et « wavey » (oie). Un segment caractéristique de cette classe : « **wayveys**, geese, ducks, **loons**, all kinds of **birds**, he looked at them. My goodness, he thought, I wonder what I

should do to kill these creatures, he thought ». Le fait que l'on retrouve également le mot « cook » dans cette classe suggère que les animaux nommés dans les contes sont des animaux consommés par les Cris. Les animaux ont une importance particulière pour les Cris notamment ceux mentionnés dans cette classe. Dans les contes, il y a souvent une personnification des animaux pour démontrer leur égalité avec les humains. Pour les Cris toute personne incluant humain, animal, esprit fait partie d'un réseau interrelié, connecté dans toute action. Ainsi, dans les contes, les relations entre animal et humain sont représentées le plus souvent par l'animal qui devient un soutien de l'humain l'aidant à régler ses problèmes (Feit 2000 :25). L'ours qui fait partie de cette classe est profondément respecté par les Cris, car d'après eux, l'ours peut comprendre le discours humain (Feit, 2000). L'hiver, quand les chasseurs cris trouvent la tanière d'un ours, ils lui manifestent du respect avant de le réveiller pour le faire sortir de sa tanière. Pour les cris « Le fait qu'un animal si fort et potentiellement si dangereux se réveille d'une manière relativement calme signifie qu'il est conscient de la requête respectueuse du chasseur et qu'il consent à s'offrir » (Feit, 2000). Les mots « wavey », l'oie des neiges ainsi que « loon » huard, se retrouvent également dans cette classe vu l'importance de la chasse à l'oie et du huard dans la culture crie. Dans cette classe, on retrouve également le mot « Weesakechahk » qui est fortement corrélé aux termes qui désignent les oiseaux, à la fois le « gray jay » ou le Whisky Jack, oiseaux qui ont inspiré le nom de ce personnage mythique filou, qui comme les geais gris joue des tours pour voler sa nourriture.

2.5. La classe 5 : Représentations de « sister »

La classe 5 (en mauve, 13.9% des segments) quant à elle, représente l'univers de « sister » et son équivalent « shishter » en référence à la sœur du personnage Chahkabesh qui lui fait la morale la plupart du temps. On raconte comment Chahkabesh, un personnage mythique cri a emprisonné (« snared ») le soleil (« sun ») par mégarde quand il cherchait la piste (« track ») de celui qui avait créé un sentier. Ainsi, l'aube (« dawn ») ne venait jamais puisque le soleil était emprisonné par le piège. Un segment caractéristique de cette classe « he said to his older **sister**, I saw **tracks** of somebody there you know big **sister**, but when he had seen the **tracks** of that person, she said to him, it's the **sun** ». Le mot « moon » fait référence à un autre conte où Chahkabesh est mis de l'avant. Selon Lefebvre (1974) Chahkabesh est un autre personnage mythique connu à travers les communautés Cries, un héros ayant des pouvoirs chamaniques et la capacité de changer de taille à volonté. Cependant, étant le cadet, ce sera sa sœur plus âgée qui endossera le rôle de sauveur tout au long du récit (Lefebvre, 1974). La présence de « moon » fait référence à la fin du récit où le héros se réfugie dans la lune, malgré les mises en garde de sa sœur : «His older **sister** said to him:'' Do not look at the **moon**,

he'll pull you up » d'où l'attraction entre le mot « moon » et « sister » dans cette classe. La figure de la sœur est importante dans la culture crie, car c'est aux enfants plus âgés que la mère lègue la responsabilité des plus jeunes.

3. Conclusion

En somme, notre choix de la méthode Reinert a bien favorisé le repérage des représentations et conceptions du monde véhiculées dans ces contes ancestraux Cris. C'est par un va et vient constant entre les mots des classes, la lecture du texte que nous avons pu interpréter ces univers de discours dégagés. Par l'entremise de cinq mondes lexicaux, l'analyse révèle la mise en discours de mythes fondateurs, d'une philosophie de vie holistique du peuple cri. Sur une toile de fond de leurs perceptions des relations familiales et des rôles distincts des hommes (détenteurs et responsables de la transmission aux générations suivantes des connaissances relatives à la chasse et à la pêche), des femmes (le feu central du foyer autour duquel gravite la famille, responsable des activités dans le campement) et de la place importante qu'occupe la sœur ainée dans la société crie (qui prend en charge les plus jeunes) se greffe leur vision égalitaire et circulaire des rapports qu'ils entretiennent entre eux, mais aussi avec les animaux, la nature et l'esprit des anciens. En outre, une place primordiale est accordée à l'eau, à la pêche et l'environnement. D'ailleurs, leur vision de la chasse révèle la relation spéciale qu'ils entretiennent avec les animaux qu'ils tuent suite à une demande respectueuse de se sacrifier pour eux, notamment l'ours, mais aussi les oies. Se greffent aussi sur l'univers familial, leurs méthodes éducatives en symbiose avec la nature et dont ils s'inspirent pour créer les personnages mythiques qui avec humour font la morale éduquant subtilement à la paix, à la régulation des relations interpersonnelles, à l'harmonie et le respect entre humains, animaux et ressources naturelles. En cette période historique importante de décolonisation, de réconciliation et de réparation des torts faits aux Autochtones du Canada, ce travail de mise en lumière valorisant leur philosophie de vie prend tout son sens.

Références

Aubry M. et Lievre P. (2016). *Project Management in Extreme Situations : Lessons from Polar Expeditions, Military and Rescue Operations, and Wilderness Exploration*. CRC Press.

Berkes F. (1979). An Investigation of Cree Indian Domestic Fisheries in Northern Quebec. *Arctic*, *32*(1): 46-70.

Davis M. (2017). *Daughters, sisters, mothers & wives*. Ontario Native Literacy Coalition.

Descola P. (2008). Claude Lévi-Strauss, une présentation. *La lettre du Collège de France*, *Hors-série2:* 4-8.

Ellis C. D. et Yereniuk R. (2014). *Cree Legends and Narratives : From the West Coast of James Bay / âtalôhkâna nêsta tipâcimôwina*. University of Manitoba Press.

Feit H. A. (2000). Les animaux comme partenaires de chasse. Réciprocité chez les Cris de la baie James Traduit de l'anglais par Jacques Jeudy. *Terrain. Anthropologie & sciences humaines*, (34): 123-142.
Grand Council of the Crees. (2019). Joyeuse chasse à l'oie 2019. Consulté 18 novembre 2019, à l'adresse *The Grand Council of the Crees*, website: https://www.cngov.ca/fr/goose-break-2019/
Jodelet D. (1997). Les représentations sociales. Dans Jodelet ed. *Les représentations sociales* (5e ed.). Presses Universitaires de France.
Lefebvre M. (1971). *Tshakapesh: Récits montagnais-naskapi*.Québec:Ministère des affaires culturelles
Ohmagari K. et Berkes F. (1997). Transmission of Indigenous Knowledge and Bush Skills among the Western James Bay Cree Women of Subarctic Canada. *Human Ecology*, *25*(2): 197-222.
Preston R. J. (2012). Cri. Consulté 1 mars 2018, à l'adresse *The Canadian Encyclopedia* website: http://www.encyclopediecanadienne.ca/fr/article/cris/
Ratinaud P. (2009). *Iramuteq : interface de R pour les analyses multidimensionnelles de textes et de questionnaires*. http://www.iramuteq.org.
Reinert M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application. *Bulletin de Méthodologie Sociologique*, vol. *26*(1): 24-54.
Reinert M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, vol. *66*(1): 5-39.
Royer M.-J. S. (2015). *Climate, Environment and Cree Observations: James Bay Territory, Canada*. Springer.
Sayles J. et Mulrennan M. (2010). Securing a Future : Cree Hunters' Resistance and Flexibility to Environmental Changes, Wemindji, James Bay. *Ecology and Society*, *15*(4): 22.

Who is Guilty and Who is Responsible in the Croatian Parliament: A Linguistic Approach

Kristina Kocijan, Krešimir Šojat

University of Zagreb – krkocijan@ffzg.hr ksojat@ffzg.hr

Abstract
From a psychological perspective, there is a difference between the terms *kriv* 'guilty' and *odgovoran* 'responsible' in that the former term conveys negative emotion and supports a predominantly passive state of being, while the latter conveys positive emotion and encourages (endorses) more active and stimulating circumstances. One's choice of term may greatly influence the communication process. It may, on one hand, bring further disruption between the sender of the communication and its receivers or, on the other hand, help them find the same path to a recovery process. This is even more important when the senders and receivers of the communication are representatives of a larger group of people, as is the case with Members of Parliament. The aim of this study is to gain insight about the political discourse in the Croatian Parliament regarding the usage of the concepts of 'guilt' vs. 'responsibility', particularly whether their usage is dependent on the speaker's gender and party preference. Therefore, we rely on a corpus with over 127 million tokens. Both qualitative and quantitative analysis are facilitated with a syntactic grammar designed within the NooJ NLP environment.

Keywords: political discourse; PDA; digital humanities; syntactic grammar; Croatian; NooJ

1. Introduction

It is quite natural that we adjust our linguistic choices depending on what we are trying to achieve with them. Curbelo (2017) discusses numerous examples where the choice of words made a difference in the political discourse and further stresses the importance of paying special attention to these choices since, as we learn from Hermann, planned speeches make language "more reflective of what the politician wants others to perceive" (Hermann, 2008:156). Furthermore, Curberlo (2017:5) stresses Bourdieu's notion that "political agents might use language strategically to maximize their symbolic efficacy." In addition, as noted by Akstinaite (2018:68), language has "a unique ability to reflect an individual's personality" and "to unveil one's thinking style". This belief is also supported by research in psychology (Tausczik and Pennebaker, 2010). Thus, we expect to find distinctive language patterns in the analysis of political discourse (cf.

Curbelo, 2017) focusing mainly on the usage of two terms: *kriv* and *odgovoran.* Furthermore, we expect that these patterns will differ depending on an MP's gender, but also her/his party orientation.

Although their meanings are quite clear when translated into English ('guilty' and 'responsible',[1] respectively) these terms are sometimes used interchangeably in Croatian. Recent research in the domain of psychology regarding the communication process (Arambašić, 2021) has linked negative emotion to the term *kriv* ('guilty') and positive emotion to the term *odgovoran* ('responsible'). In order to test how this theory reflects upon the speeches of the Members of the Croatian Parliament (MPs), we have designed a syntactic grammar that detects our terms of interest and their near surroundings. In the results obtained, we have observed several patterns of usage of the aforementioned two terms. It is important to note here that the purpose of this research is not to search for people/groups associated with the usage of the terms *kriv* or *odgovoran*, but rather to detect language patterns that occur with these two terms. The remainder of this article takes the following structure: Section 2 discusses related work on content analysis of political texts; Section 3 and 4, respectively, provide a description of our data set and the unsupervised algorithm that was used to detect different language patterns; Section 5 reports our results. We conclude with a short discussion of our findings.

2. Previous work

The language of the political arena has been of interest to many researchers in different fields (historians, biographers, psychologists, political scientists, linguists, and sociologists, *inter alia*) (cf. Kocijan and Šojat, 2021). Due to the different nature of their research questions, a variety of algorithms is proposed to help us better understand the world via political text analysis. And while, not so long ago, text analysis was done with the help of human experts only (Tausczik and Pennebaker, 2010), today we can use the power of a computer to perform the same job with more speed and over a larger data set. Regarding Croatian political discourse, Korenčić *et al.* (2016) used semi-supervised topic modeling methodology for a preliminary content analysis of media topics during the 2015 parliamentary election in Croatia. The most recently proposed semantically aware, unsupervised algorithm is SemScale, which is used for political text scaling (Nanni *et al.*, 2021) (i.e., determining Left-Right political

[1] 'responsibility' is 85/237 linguistic marker on the WEKA (Waikato Environment for Knowledge Analysis) list of the most promising features predicting hubristic speech with the information gain score set to 0.00914 – the higher the score, the higher the predictive power (Akstinaite, 2018:292).

ideology) with impressive accuracy on five different languages spoken in the European Parliament (EP). This algorithm outperforms Wordfish, an unsupervised model proposed by Proksch and Slapin (2009). Unlike these two models, which are designed to determine the political ideology based on the language used, we will take a different approach and see how language is used within a certain political ideology.

3. Corpus

The Open Government Data policy makes Croatian Parliamentary debates freely available on their web site[2] in PDF format. The files are transcripts that date back to 2003, from the fifth legislative term of the Croatian Parliament, and include information on the *legislative term*, *meeting, date, agenda item number, the full title of an agenda item, MP's name and party affiliation,* and the *text of his/her speech.* Similar data exists for other languages, such as the Bundestag corpus of German parliament speech (Biessmann *et al.*, 2016), and Hansard corpora of English spoken in the Canadian, Australian, and New Zealand parliaments, as well as the EuroParl corpus, which, in addition to English, provides parallel data on the European Parliament in 24 official languages (Proksch and Slapin, 2009; Nanni et al., 2021, *inter alia*). For the purposes of our experiment, after downloading the source files, we converted them into excel tables in order to build two specialized data sets: gender oriented and party oriented, each falling within the domain of political discourse as defined by Van Dijk (1997). Due to the nature of our research, we opted to aggregate individual speeches of MPs to the national party level, but also to the level of political ideology (Left-Right-Center-Regional[3]) (on the positive and negative aspects of such an approach, see Proksch and Slapin, 2009). This allowed us to include all available speeches regardless the text length per specific speaker or party. Furthermore, gender is additionally marked for each speaker to accommodate for the language analysis at the gender level. Finally, each entry is given a numeric ID marker that will help us retain the original order of speeches and also serve as a reference point. At this time, there was no need to make any other pre-processing since we had opted for the NooJ NLP environment (Silberztein, 2016) to automatically detect and tag the desired language models describing 'guilt' and 'responsibility', as will be described in the following section. In spite of the fact that language is a multi-layered system, we did not consider any non-linguistic dimensions (e.g.,

[2] Hrvatski Sabor, informacijsko-dokumentacijska služba - https://edoc.sabor.hr/

[3] The size of the Regional sub-corpus was too small and thus uncomparable to other three sub-corpora. For this reason, it was not included in the further analysis at the level of party orientation. However, they were included in the gender-oriented corpora.

gestures, pauses, facial expressions) for this specific research. The full Croatian Parliament Corpus - CroParlCor contains 127,263,579 tokens and is constructed out of six files, each consisting of data for one legislative term from 2003 up to 2021. Each legislative term has a different length (in months) but also in size (number of tokens)[4]. Although Biessmann *et al*. (2016) characterize parliament debates as the best source for automated political text analysis, from a linguistic perspective, this corpus is not representative of a political corpus *per se*, since the language MPs use varies from setting to setting and is dependent on the format of the political context (interviews, media coverage, party manifestos, etc.). However, CroParlCor is representative from the situational perspective, since we define our register as *transcriptions of all MPs' speeches in the Croatian Parliament* from the end of 2003 to the end of 2020 (cf. Bieber, 1993 for further analysis of linguistic vs situational perspectives of a corpus). Since the full CroParlCor corpus was constructed, for the purposes of this research, we have restructured it two times (models A and B) in the following manner:

A CroParlCor-Gender driven corpus (124,659,320 tokens)
  - A.1 Speeches by female MPs (25,753,090 tokens) and male MPs (98,906,230 tokens) are divided into two separate sub-corpora
  - A.2 Within each sub-corpus, speeches belonging to the same calendar year from 2003 to 2021 are saved as individual texts[5] (23 texts per sub-corpus), allowing us also to observe changes through time

B CroParlCor-Party driven corpus (101,313,884 number of tokens)
  - B.1 Speeches by each MPs party orientation are divided into three separate sub-corpora: Left-oriented parties (36,021,586 tokens), Center-oriented parties (26,836,080 tokens), Right-oriented parties (38,456,218tokens)
  - B.2 Within each sub-corpora, speeches from the same year are saved as individual texts (23 texts per sub-corpora) allowing also to observe the changes through time.

The difference in the number of tokens between the full corpus and corpus A (gender-oriented) vs corpus B (party-oriented) is accounted for by the difference between the metadata included in them: unlike the full CroParlCor, which includes all the information (as described earlier), corpora A and B include only

[4] Term V (22.12.03 – 11.1.08) – 20,835,650 tokens; Term VI (11.1.08 – 22.12.11) – 17,646,746 tokens; Term VII (22.12.11 – 28.12.15) – 26,651,367 tokens; Term VIII (28.12.15 – 14.10.16) – 2,319,836 tokens; Term IX (14.10.16 – 22.7.20) – 42,954,028 tokens; term X (22.7.20 – 31.12.21) – 16,855,952 tokens

[5] This is true of all the years except 2011, 2015, 2016, and 2020, since they were partially in 2 consecutive legislative terms. For these years, two files were formed – one per term.

the ID and the speech itself. Further differences between corpus A and corpus B are due to the number of speeches included in each.

4. Searching for the language models

The unsupervised algorithm was implemented in the **first phase** of the project aiming to detect all occurrences of *kriv* and *odgovoran*. Our analysis of the data revealed several models of their usage that helped us detect their semantic variations. In the **second phase,** we augmented the algorithm with additional notations in the transducer, marking the output with tags for the type and its subtype <KRIV+KRIV | POGRESAN | POGRESNO | NEDUZAN | ZAO_MI_JE> and <ODGOV+ODGOVORAN | ODGOVORNO | ODGOVORNOST>. We also tagged the verbs, NPs, and PPs that are found in the vicinity of the term such that the verbs and nouns are lemmatized (i.e., written in their canonical forms). Their normalization allowed us to reduce overall word variations giving us more precise measures. Since we were not interested in all meanings of the word *kriv* (that is, we were not interested in its interpretations like 'wrong', 'incorrect', or 'curved') we removed these outputs from our results. The following sections provide an analysis of our findings.

5. Results

5.1. Results for the gender-driven corpora

Results for the gender-driven corpora show two distinctive patterns in the use of the term *kriv* (Figure 1. – red line) and *odgovoran* (Figure 1 – green line) through time (from 2003 up to 2022). It is quite clear that these terms are used differently by the two genders, and that female MPs preferred talking about responsibilities until 2013, with the highest peak being in 2012, when the standard score (z-score) was statistically significantly higher than expected (z-score = 2.364). The terms that co-occur with the term 'responsible' in 2012 can be divided into three modules: those that co-occur with 'responsible' in the form of an **adverb** (*odgovorno,R* 'responsibly') denoting responsible behavior (*zaokružiti* 'to circle'*, pristupiti* 'to approach' *etc.*), in the form of an **adjective** (*odgovoran,A* 'responsible') with 4 subcategories detected (Government: *oporba* 'opposition'*, država* 'state' *etc.;* Institutions: *članica* 'member'*, fakultet* 'faculty' *etc.*; People: *čovjek* 'man'*, društvo* 'society' *etc.;* Things: *karika* 'link'*, mjesto* 'place' *etc.*), and in the form of a **noun** (*odgovornost,N* 'responsibility'). In the third module, we detected a set of **verbs** that co-occur with this noun (*amnestirati od* 'to grant amnesty from'*, biti oslobođen* 'to be released' *etc.*) but also types of responsibility as modified by **adjectives** (*automobilska* 'automotive'*, društvena* 'social'*, financijska* 'financial' *etc.*), by **prepositional phrases** (*za praćenje procesa* 'for process monitoring'*, za procjenu* 'for assessment' *etc.*), and by **noun phrases** (*pravne osobe* 'legal entities'*, roditelja* 'parents'*, suca* 'judge' *etc.*).

Since 2014, they have twice dropped below the threshold (from -2 to 2) for the usage of *odgovoran* (z-score[2018] = -2.006; z-score[2021] = -2.606) and once for the usage of *kriv* (z-score[2018] = -2.454. In the male corpus, we observe statistically significant results in 2013, when they use the term *kriv* more often (z-score = 3.000), and again in 2018, when the standard score for both terms drops below the threshold (z-score[*kriv*] = -2.028 and z-score[*odgovoran*] = -2.191) and continues to drop for the term *odgovoran* in the following year (z-score [*odgovoran*] = -2.267). Among the results for *kriv* in 2013, we detected six subcategories of words and phrases that co-occur with the term *kriv* as an **adjective** in the context of 'being guilty' (Government: *država* 'state', *Vlada* 'government', *bivša Vlada* 'former government' *etc.*; Institutions: *administracija* 'administration', *agencija* 'agency', *bolnica* 'hospital' *etc.*; People: *djeca* 'children', *ljudi* 'men, *narod* 'people' *etc.*; Things: *gospodarski status* 'economic status', *godina* 'year', *konstelacija* 'constellation' *etc.*; Professions: *liječnici* 'doctors', *nezaposleni* 'unemployed', *poduzetnik* 'entrepreneur' *etc.*; and Pronouns: *ja* 'I', *mi* 'we', *nitko* 'no one' *etc.*).

5.2. Results for the party-driven corpora

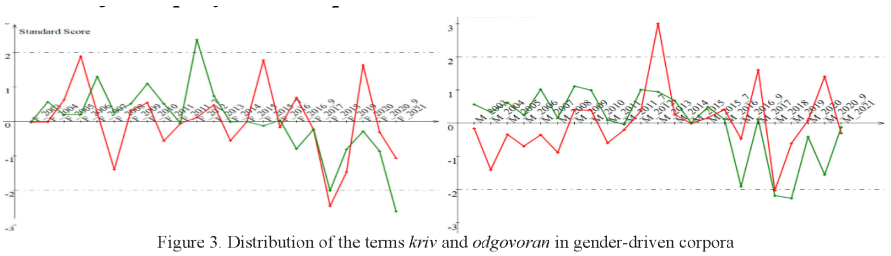


Figure 3. Distribution of the terms *kriv* and *odgovoran* in gender-driven corpora (left graph for female MPs, right graph for male MPs)

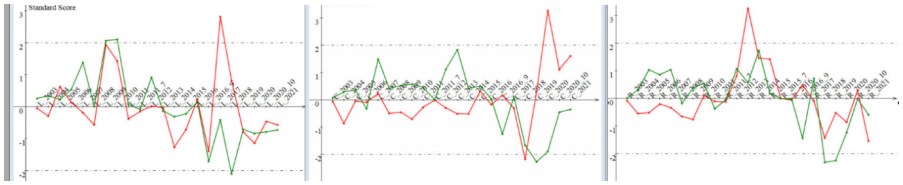


Figure 1. Distribution of guilty (red line) and responsible (green line) within three political orientations:
Left parties (far left graph) – Center parties (middle graph) – Right parties (far right graph)

In line with our main hypothesis regarding the difference in patterns detected by MPs with different political orientations in a party-driven corpus, our preliminary

expectations were correct. All three sub-corpora manifest different usage patterns with three distinctive peeks concerning use of the term *kriv* occurring in different periods of time (Figure 2). Thus, the graphs in figure 2 show that left-oriented parties show significantly higher usage of that term in 2017 (z-score = 2.818), center-oriented parties in 2020 (z-score = 3.264), and right-oriented parties in 2013 (z-score = 3.262). Two years in a row, in 2009 and 2010, left-oriented parties showed statistically significant results for the use of term *odgovoran* (z-score[2009] = 2.060; z-score[2010] = 2.105), and in 2018, they significantly failed to use that term (z-score[2018] = -2.100). The usage of *odgovoran* for the other two party orientations shows no high positive results for standard score but rather negatively marked statistically significant results, once for center-oriented parties (z[2019] = -2.272) and twice for right-oriented parties (z[2018] = -2.314; z[2019] = -2.249).

6. Conclusion

This research opens the door to other at-a-distance context analysis of political texts. The results can also be used within the Voting Advice Applications (VAAs) or the more recent Conversational Agent VAA's (CAVAA's) that are being used in modern electoral campaigning as on-line applications that give voters voting advice by providing a candidate's or party's standpoint on different topics (Kamoen and Liebrecht, 2022). Although this seems to be a popular method used across many European countries, to our knowledge, no VAA or CAVAA has been built for Croatian voters so far. In our future work, we plan to include more precise semantic analysis of the terms we discuss here, namely *krivnja* and *odgovornost*. We intend to focus on the semantic structure of these terms, i.e. on senses that can be determined in their overall meaning, and to examine whether they can be incorporated in more elaborate and precise language models. Our plan is also to introduce potential semantic distinctions into our tagging scheme and to examine their collocational spans along the principles described above. When combined primarily with data on the party and gender of speakers, but potentially also on the topics discussed, such models could provide an excellent basis for the analysis of political attitudes and communication strategies used in parliamentary debates.

References

Akstinaite V. (2018). Use of Linguistic Markers in the Identification and Analysis of Chief Executives' Hubris. PhD Thesis, Faculty of Arts and Social Sciences, University of Surrey.

Arambašić L. (2021). *Svemoć i nemoć komunikacijskog procesa*. Naklada Slap.

Biber D. (1993). Representativeness in Corpus Design. In Literary and Linguistic Computing, vol. 8 (4): 243-257.

Biessmann F., Lehmann P., Kirsch D. and Schelter S. (2016). Predicting political party

affiliation from text. In Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), Dubrovnik, Croatia, pp. 14-19.
Curbelo A. A. (2017). Analysing the (Ab)use of Language in Politics: The Case of Donald Trump. Working Paper n. 2. University of Bristol: SPAIS.
Hermann M. G. (2008). Content Analysis. In Klotz A., Prakash D. (eds) *Qualitative Methods in International Relations*. Research Methods Series. Palgrave Macmillan, London, pp. 151-167.
Kamoen, N. and Liebrecht C. (2022). I need a CAVAA. How Conversational Agent – Voting Advice Applications (CAVAA) affect users' political knowledge and tool experience. *Frontiers in Artificial Intelligence.* doi:10.3389/frai.2022.835505
Kocijan, K. and Šojat, K. (2021). Negation Usage in the Croatian Parliament. In Bigey, M., Richeton, A., Silberztein, M. and Thomas, I. (eds), *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, Switzerland: Springer, Cham, 2021. pp. 101-113.
Korenčić D., Grbeša-Zenzerović M. and Šnajder J. (2016). Topics and their Salience in the 2015 Parliamentary Election in Croatia: A Topic Model based Analysis of the Media Agenda. In Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), Dubrovnik, Croatia, pp. 48-54.
Nanni F., Glavaš G., Rehbein I., Ponzetto S.P. and Stuckenschmidt H. (2021). Political Text Scaling Meets Computational Semantics. Available at: https://arxiv.org/pdf/1904.06217.pdf
Proksch S. O., and SLAPIN J. B. (2010). Position Taking in European Parliament Speeches. *British Journal of Political Science, 40(3)*, pp. 587–611. http://www.jstor.org/stable/40930601
Silberztein, M. (2016). Formalizing Natural Languages: The NooJ Approach, Cognitive science series, Wiley-ISTE, London, UK.
Tausczik Y. R. and Pennebaker J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. In *Journal of Language and Social Psychology 29(1)*, pp. 24-54.
Van Dijk, T. A. (1997). What is political discourse analysis? In: Bloomaert, J., Bulcaen, Ch. (eds.) Political Linguistics, Amsterdam, pp. 11-52.

Dynamic and content of COVID-19 related tweets during early stage of pandemic

Tomasz Kruszewski, Joanna Michalak

Nicolaus Copernicus University in Torun – tomkrus@umk.pl
joanna.michalak@umk.pl

Abstract
On March 11, 2020, the COVID-19 was identified by WHO as a global pandemic. At the beginning of which the uncertainty influenced the emotions of investors. This article aims to describe dynamic and content of investors COVID-19 related tweets in March 2020 in terms of Ekman classification of emotions. To describe the content of tweets n-gram and KMeans procedure were used as well as lexicon-based sentiment analysis (NRC lexicon). To describe the dynamics ARIMA(p,d,q) model was used.

Keywords: behavioural economy, emotions, covid-19, twitter, sentiment analysis, business psychology

1. Introduction

Pandemic cause main three group of problems, (1) public health (medical shocks), (2) economic (demand and supply shocks) and (3) psychological problems (expectations, attitudes, and perception). These groups influence each other and are a source of uncertainty (Michalak & Kruszewski, 2021). Pandemic uncertainty affects society as a whole and certain groups, such as individual investors. It throws people out of the framework of typical activity (Chawla et al, 2021). Pandemic stress affects the neurochemical processes in human body and interpretation of received signals. Uncertainty may increase or decrease because of stream of information. Emotions caused by information influences rationality, cause cognitive biases or short path in decision making, heuristics (Galea et al, 2020). If uncertainty cooperate with intense emotions in large group of investors cause contagion of market players and can reverse market trend, lead to big drops, or increase volatility. The proxy variable of the aggregate emotions of investors are semantically processed tweets. By analysing their causal relationship with stock market indicators, we may understand the behavioral drivers of price movements. The aim of this article is to describe dynamic and content of COVID-19 related tweets in March 2020.

2. Literature review

There was a monotonic upward trend in the way users interact with COVID-19 related information after January 20, 2020 (Michalak & Kruszewski, 2021). Information volatility of Twitter and stock market depended on speed of virus reproduction. Twitter showed local patterns caused by higher spread of virus and has predicted stock market indicators (Huang et al, 2020; Albulescu, 2020; Alfaro et al, 2020; Baek et al, 2020). Economic leaders were also of great importance, 203 COVID-19 related tweets sent by G7 leaders went viral. Any of these tweets had potential to impact stock market. President Donald Trump was one of the leaders with biggest interactions in social media during COVID-19. Some studies shown conflicting emotional signals as reaction to new reality. However, many indicate that fear and anger were one of the most common emotional responses to pandemic (Ahorsu et al, 2020; Kruszewski & Michalak, 2020).

3. Noise investors reaction

3.1. Research methodology

We tested 3 hypothesis which allows us to represent the content and dynamic of tweets. Hypothesis 1 lead to describe emotional profile of investors in March 2020. Hypothesis 2 lead to answer if investors habituate to pandemic, which can be represented by negative trend. According to the definition of emotions, we assume that explored indicators should be characterized by short memory of the process (H3). In other words, y_t is correlated with its past values y_{t-i} in short time rather that in long time. If the procedure shows a short memory of the process, then we can assume that indicators are emotions by definitions. If process is characterized by long memory the effects can be cumulative and form some mood. The mood can influence investors perception in long run if user stay in the same information bubble.

Process of hypothesis verification included following steps:

- H1: The dominant emotion among investors in March 2020 was fear
  a) Pre-processing/ removing: numbers, URL, hashtag, cashtag, username, emoticons, punctuations and stop list.
  b) Pre-processing/ replace: contractions, multi punctuation marks and elongated words.
  c) Analysis with lexicon based approach, NRC (Michalak, 2020).
- H2: Volume of COVID-19 related tweets is characterized by negative trend
  a) Analysis of regression with *time* and $time^2$ variable representing linear trend and square trend

$$Y_t = \beta_0 + \alpha_1 time + \varepsilon_t \quad (1)$$

$$Y_t = \beta_0 + \alpha_1 time + \alpha_2 time^2 + \varepsilon_t \quad (2)$$

- H3: Time series of emotions is characterized by short-term memory of the process
  a) To identify ARIMA*(p,d,q)* model we used Box & Jenkins procedure (Kufel, 2020).

$$(1-u)^d Y_t = \alpha + \delta y_{t-l} + \delta y_{t-2} + \ldots + \delta p y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_1 \varepsilon_{t-2} + \ldots + \Theta_q \varepsilon_{t-q}, \quad (3)$$

  where u is the time-shift operator $u^d Y_t = Y_{t-d}$, p parameter is lag order in AR(p) component and q parameter is lag order in MA(q) component.

We also used Pearson correlation indicator to determine the correlation between two pairs: (1) stock market and COVID-19 related tweets, (2) COVID-19 related tweets and pandemic basic indicators, such as: new cases, total deaths, new deaths. We ask if there was any premise (beside evidence from other papers) that our Twitter indicators are connected to stock market. Next, to in deep look into selected emotion, we apply Kmeans unsupervised algorithm. This allows us to summarize selected emotion by topics that are distributed in them.
Twitter data has been collected by Twitter API with Python software. Stock market data comes from Yahoo Finance. Pandemic indicators come from Johns Hopkins University Centre for System Science and Engineering. The JHU CCSE dataset is available in Dbnomics, which enables automatic data download by GRETL software (Kufel, 2020). Content analysis was performed in Python and dynamic analysis was performed in GRETL software.

3.2. Results

The sum of tweets which contains the stock exchange cashtag (March 2020) was 649,548 with a daily average of 20,953 tweets, maximum of 49,335 and minimum of 7,764. The subset of tweets related to pandemic, selected as those containing the words from the list [COVID-19, covid, coronavirus, pandemic], includes 45,837 tweets. The daily mean was 1,478 with maximum value of 4,649 tweets and minimum value of 308. 7,08% of all tweets was relate to a pandemic. Figure 1b shown tweets volume with daily time window. The lack of zero volume days shows importance of pandemic related events.

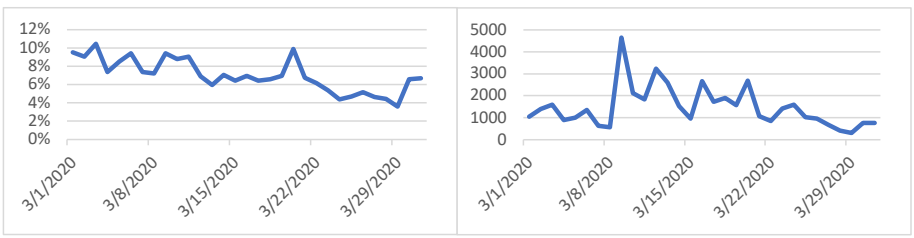


Figure 1. [a] Percentage of tweet containing the 'COVID-19 & stock market' in March 2020 [b] Daily number of tweets posted in March 2020 with keyword 'COVID-19 & stock market'

Pearson correlation indicator showed negative correlation between Twitter volume and set of pandemic indicators. Beside no interpretative value we assume that the negative correlation is caused by opposite, deterministic trend in time series. After detrending data these indicators become not statistically significant. We find statistically significant correlation between volume of tweets and stock market measured by volume, $\rho = 0{,}49$. We assume that there is an indirect relation between pandemic and tweets.

We looked for periodicity by including set of 0-1 variables in regression for hourly time window and did not find patterns of deterministic periodicity. This indicate that informational landscape is driven by political decisions and decision-makers moves, not by human biological clock, like time of the day (activity time/sleep time). Policy decisions are determined by local reproduction of virus indicators and expectations. For example, outlier from March 9, 2020, can be related to expectations of WHO decisions according to dynamic growth of infections. In fact, March 11, 2020, was the day of WHO's classification of world events as a pandemic. Next volume tick occurs on March 16 when S&P500 and NASDAQ Composite fell 12%, which where one of the highest drops nowadays.

Results of trend equation estimation shows that there is no statistically significant trend in volume of tweets time series (figure 1b). On the other hand, we counted the percentage of COVID-19 related tweets in set of all $stock_market related tweet dataset (figure 1a). We treated this series as an indicator of the intensity of the discussion. This series is characterized with statistically significant downtrend with non-intraday periodicity patterns. From psychology point of view and Taleb's theory of uncertainty point of view downtrend is natural due to the adaptation to the new reality. Ramlli and Wagner (2020) provide a time analysis of intensity of pandemic with 3 stages of initial pandemic phase proposition: incubation (January20 to January 17), epidemic (January 20 to February 21) and fever (February 24 to March). In our research, the incubation phase contained a low number of tweets as well as epidemic phase. There is a significant increase of tweets volume in fever phase. After the break-point events, habituation process can be visible (with negative trend).

Set of bigrams, constructed on basis of all tweets, shows that investors focus their online conversation on technical market analysis: view odds (6337, the most important bigram in March 2020), 'moving average' (5167), 'data moving, momentum indicators' (4959;3861). The general condition of the market is assessed as negative: 'market enters downtrend' (1456 mentions) in first 4 days with more negative 'market crush' expectations in next 4 days and 'finals market crush' in March 10-13. 'Corona virus' bigram occur for first time in March 5-9, at position 25 (volume: 1008 / 14,401 all tweets).

Set of bigrams related to COVID-19 shows that investors concentrate at market risk and its evaluation ('market correction', 484/1000). Discussion displays the current events which affect market, for example 'oil prices (361 in March 5-9)' or in general 'live news'. Researchers conclude that indicator which affect investor expectations and investment decision is indicator of reproduction of COVID-19 not by mortality indicator. It is because the increase in the number of infections is associated with restrictive restriction. Our study confirms this by a large volume of 'coronavirus spread' bigram (510/1000 in first 4 days of March) and lack of 'mortality' related bigram. Compared to the dataset with all tweets, there has been

a change – from technical analysis related bigrams to emotional tone of this conversation ('fear' (123), 'panic' (123) unigrams). Other paper shows that fear and anxiety were first of emotions strongly associated with pandemic first wave. These emotions drive financial decisions and perception.

For a better understanding of the content of tweets we extended the bigram analysis with KMeans procedure. In anger emotion related tweets, the Elbow method indicates k=5 groups. We find most relevant words for each of these group, there are 3 groups which are connected to technical analysis of market. Investors search information and discuss market trends and technical indicators. High use of words '*buy-sell*' may indicate expectations to changes in the price level or/and the impact of bots. Last 2 groups contain tweets connected to (1) *market*, *crush*, *loss*, *fear*, *virus*, *bear*, *death*, *economy* and (2) *Trump*, *blame*, *president*, *virus*, *crash*, *elect*, *test*, *Obama*. Results confirms that increase in pandemic indicators, and current political actions were highly discussed. A quick review of these tweets showed us that they mainly concern on election and try to predict president decisions leading to stop spread the virus. In fear emotion related tweets, the Elbow method indicate k=4 groups. Two of these groups indicate discussion about market trends (*average*, *moving average*, *economy*, *odd*, *indicators*) and election (*Trump*, *economy*, *Biden*, *election*). Others contained words indicating an emotional response of investors: *lose*, *bad*, *fear*, *crash*, *broken*.

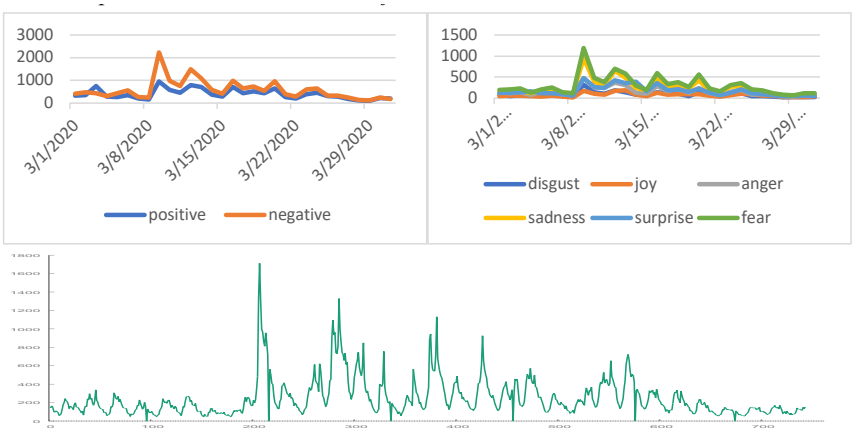


Figure 2. Dynamic of [a] positive and negative emotions in March 2020, [b] emotions according to Ekman theory of emotions, [c] fear tweets dynamic

Figures 2 sum up results of sentiment analysis. Negative, fear and sadness emotions had the highest intensity in March 2020. Most of researchers suggest that fear was driving force in stock market. The increase in the intensity of emotions is visible with intensification of coronavirus reproduction in Europe

and the response of Europeans stock markets. Increase in pandemic related tweets coincides with establishment of COVID-19 epicentre in Europe.

For both variables KPSS test and ADF test indicate stationarity of the process at I(0). For a vast majority of the estimated ARIMA*(p,0,q)* models for *p,q*-{0,1,2,3,4} the minimum value of the AIC and BIC criterion pointed on the ARIMA(1,0,0) model. An essential feature of the ARIMA (1,0,0) is modelling Y_t by the autoregressive model of order 1. The estimated equation, φ_1 satisfies the condition of AR process stationarity ($|\varphi_1| < 1$). For both models, parameter has positive value which means that ACF exponentially decays to 0 as lag increase. The intercept indicates a daily increase in news (about 250 tweets for both variables). Which can be interpreted as influence of fever period in our sample with all outliers which occur there. The AR(1) part indicates the dependence of the mean of the process on past values (1 hour). In both cases we can say that the short-memory condition of the process was met because higher lags were not statistically significant. A summary of the ARIMA(1,0,0) models is provided in table 1. Because series in their dynamic were very similar, the estimation result also indicate likelihood. The procedure of studying the autocorrelation and the structure of presented processes will be extended in the future research.

Table 1. Results of ARIMA(1,0,0) estimation for fear and anger variables

Start	End	Const.	Phi_1	p	R^2	Additional information
Fear						
01-03-2020	31-03-2020	243,461 (30,19733)	0,88862 (0,0166690)	$p_{(const.)}$=7,49e-16 $p_{(phi\_1)}$=0,000	79%	α=5%
Anger						
01-03-2020	31-03-2020	240,462 (29,983)	0,8786 (0,00266)	$p_{(const.)}$=7,59e-16 $p_{(phi\_1)}$=0,000	78%	α=5%

3.3. Conclusions

Since investors formulate their expectations about pandemic, consume and create information in social media, institutions can, with limitations, influence these expectations through social media. From evidence, presented in literature, we know that COVID-19 had an impact on the correction of expectations regarding the increase in dividends in the USA and Europe. The results of presented pilot study indicate the dominance of two emotions, which strongly influence the decision-making process of investors. After the, so-called, fever stage of first wave of pandemic discussion shows a tendency to mute (measure by percentage

of COVID-19 related tweets to all of tweets send by investors). In future studies, the research period will be extended. Here it was chosen due to the black swan theory, which indicates that the most important reactions should be visible in the first phase of the event.

References

Ahorsu D. K., Lin C. Y., Imani V., Saffari M., Griffiths M. D. and Pakpour A. H. (2020). The fear of COVID-19 scale: development and initial validation. *International Journal of Mental Health and Addiction*. World Wide Web: https://link.springer.com/content/pdf/10.1007/s11469-020-00270-8.pdf [access: 15.03.2022].

Albulescau C. T. (2021). COVID-19 and the Unites States financial market volatility. *Finance Research Letters*, vol 38. World Wide Web: https://www.sciencedirect.com/science/article/pii/S1544612320303202 [access: 6.05.2022].

Alfaro L., Chari A., Greenland A. and Schott P. (2020). Aggregate and firm-level stock returns during pandemics, in real time. *NBER Working Paper*, no. 26950. World Wide Web: https://www.hbs.edu/faculty/Pages/item.aspx?num=57930 [access: 6.05.2022].

Baek S., Mohanty S. K. and Glambosky M. (2020). COVID-19 and stock market volatility: and industry level analysis. *Finance Research Letters*, vol. 37. World Wide Web: https://www.sciencedirect.com/science/article/pii/S1544612320311843?via%3Dihub [access: 6.05.2022].

Baker S. R., Bloom N., Davis S. J., Kost K. J., Sammon M. C. and Viratyosin T. (2020). The unprecedented stock market impact of COVID-19. *NBER Working Paper Series*, no. 26945. World Wide Web: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7467874 [access: 6.05.2022].

Chawla C., Mangaliso M., Knipes B. and Gauthier J. (2012). Antecedents and implications of uncertainty in management: a historical perspective. *Journal of Management History*, vol. 18(2): 200-218.

Galea S., Merchant R. M. and Lurie N. (2020). The mental health consequences of COVID-19 and physical distancing: the need for prevention and early intervention. *JAMA Internal Medicine*, vol. 18(6): 817-818.

Huang J., Wang H., Xiong H., Fan M., Zhuo A., Li Y. and Duo D. (2020). Quantifying the economic impact of COVID-19 in Mainland China using human mobility data. *Economics. General Economics*. World Wide Web: https://arxiv.org/abs/2005.03010 [access: 15.03.2022].

Kruszewski T. and Michalak J. (2020). Irritability and anger in COVID-19 related tweets: its psychological content and dynamic (introduction to research of black swan event impact on stock market). In Soliman, K. editor, *36th International Business Information Management Association Conference (IBIMA), Granada, Spain, 4-5 November 2020*. IBIMA.

Kufel T. (2020). Arima-based forecasting of the dynamics of confirmed covid-19 cases for selected European countries. *Equilibrium: Quarterly Journal of Economics and Economic Policy*, vol. 15(2): 181–204.

Michalak J. (2020). Does pre-processing affect the correlation indicator between twitter message volume and stock market trading volume?. *Ekonomia i Prawo*, vol. 19(4): 739-755.
Michalak J. and Kruszewski T. (2021). *Pandemia a reakcje inwestorów. Analiza big data komunikatów z Twittera wobec zjawiska niepewności*. TNOiK 'Dom Organizatora'.

The affecting meaning of words: multidimensional analysis of emotional communication

Daniela Laricchiuta[1], Andrea Termine[1], Carlo Fabrizio[1], Noemi Passarello[2], Francesca Greco[3], Fabrizio Piras[1], Eleonora Picerni[1], Debora Cutuli[3], Andrea Marini[4], Laura Mandolesi[2], Gianfranco Spalletta[1], Laura Petrosini[1]
[1]IRCSS Fondazione Santa Lucia – daniela.laricchiuta@gmail.com andreatermine544@gmail.com carlo.fabrizio217@gmail.com f.piras@hsantalucia.it eleonora.picerni@gmail.com g.spalletta@hsantalucia.it laura.petrosini@uniroma1.it
[2]University of Naples Federico II – noemi.passarello@gmail.com laura.mandolesi@unina.it
[3]Sapienza University of Rome– francesca.greco@uniroma1.it debora.cutuli@uniroma1.it
[4]University of Udine, Italy – andrea.marini@uniud.it

Abstract
The abilities to be aware of one's own emotions and to translate emotional experiences into words are crucial for well-being and relationships. Notably, translating emotional experience into words is associated with language style.. This contribution analyzed healthy adults' ability to express and identify emotions by means of the Toronto Structured Interview for Alexithymia (TSIA) and to regulate emotions by means of a specific psychological assessment combining Attachment Style Questionnaire, Emotion Regulation Questionnaire, Coping Orientation to Problems Experiences, Beck Depression Inventory, and State–Trait Anxiety Inventory-2. TSIA transcripts and recordings were analyzed by Natural Language Processing and OpenSMILE API to shed light on verbal symbolic and subsymbolic features. The brain structural correlates of the capacity to translate emotional experience into words were determined through cortical thickness measures. A Machine Learning methodology proved that individuals with deficits in identifying and describing emotions had the following characteristics: 1) They were unconfident and uncomfortable with relationships, inhibited their emotional expression, and adopted avoidance coping. 2) They produced language distortions, frequently used the present tense of auxiliary verbs, and few possessive determiners. Their speech was scarcely connected and not characterized by emotional keywords; speech semantic structure was associated with bodily sphere. 3) They showed high cortical thickness at the left temporal pole and low cortical thickness at the isthmus of

the right cingulate cortex. 4) Their voice showed few variations in harmonicity, quality, or signal, whereas the intensity was irregular. Overall, the present contribution is based on cross-disciplinary approach allowing getting a whole uptake of the capacity to verbally express emotional experiences into verbal forms, along with its biological correlates.

Keywords: Alexithymia, Emotional Processing, Cortical Thickness, Natural Language Processing, Machine Learning

1. Introduction

Language provides insight into emotions, since the representations of each word we use are maintained in a mental lexicon containing information about pronunciation, syntax, general meaning, and affective connotations. The Multiple Code Theory holds that emotional information is encoded both verbally and non-verbally in symbolic and subsymbolic systems (Bucci, 2021): Namely, the verbal symbolic system refers to the capacity of language to direct and regulate action and to activate imagination and emotion, and the verbal subsymbolic system concerns aspects, such as prosody, meter, rhythm and phonemic qualities of language. While the symbolic dimension arranges the sequence of words to be comprehensible, the subsymbolic dimension influences the selection of words. A translation of emotional experience into words requires shifting from implicit (subsymbolic) to explicit (symbolic) levels, operation depending on what is called "referential process" (Bucci, 2021). Contemporary neuroscience retains that biological correlate of expression and experience of emotions are distributed across brain circuits and bodily organs (LeDoux, 2021).

Specifically, emotional expression occurs at a physiological and motor-expressive level, sometimes in an implicit way, and involves changes in visceral states and somatosensory cortical and limbic areas. As narratives inherently mirror emotional processes, it is reasonable to assume that the translation of emotional experience into words is influenced by language style. Therefore, the interpretation of verbal symbolic and subsymbolic elements of narratives revealed during an interview on the capacity to identify and describe emotions is a powerful tool to reveal how emotion is handled.

The present study utilized transcripts and recordings obtained through the Toronto Structured Interview for Alexithymia (TSIA) (Bagby et al., 2006). Attachment Style Questionnaire (ASQ), Emotion Regulation Questionnaire (ERQ), Coping Orientation to Problems Experiences (COPE), Beck Depression Inventory (BDI), and State–Trait Anxiety Inventory-2 (STAI-2) were administered to the same sample of healthy subjects performing TSIA.

After the psychological assessment, cortical thickness was measured by 3-T high-resolution structural Magnetic Resonance Imaging (MRI). TSIA transcripts and recordings were analyzed through Natural Language Processing (NLP) (Benoit

et al., 2018) techniques and OpenSMILE API (Eyben et al., 2010) to extract acoustic and semantic features.
Finally, cross-disciplinary (psychological, linguistic, acoustic, and neuroimaging) outcomes were combined through Machine Learning (ML) to provide an overview of the elements that contribute to emotional expression.

2. Materials and methods

2.1. Participants

Healthy adults (aged 21-63 years) were recruited at the IRCCS Fondazione Santa Lucia, Rome, under typical exclusion and inclusion criteria detailed elsewhere (Laricchiuta et al., 2014).

2.2 Psychological assessment

Participants underwent psychological assessment by means of TSIA, ASQ, ERQ, COPE, BDI, and STAI-2. Italian versions of the psychological scales were used.

2.3 NLP analysis

TSIA transcripts were pre-processed using Universal Dependencies (UD) Pipe (Straka et al., 2016). Linguistic profiling was carried out using Profiling-UD (Brunato et al., 2020). All features extracted in the linguistic profiling were tested by univariate correlation filtering using the Spearman method.
Semantic profiling was carried out using the $\chi 2$ test. Topics were identified using Latent Dirichlet Allocation (LDA) (Benoit et al., 2018). Sentiment Analysis was performed using the Open Polarity Enhanced Name Entity Recognition (OpeNER). Emotion Recognition Analysis was performed using the Canada National Research Council (NRC) Word–Emotion Association Lexicon. All features were compared between groups using multiple Wilcoxon tests.

2.4 MRI acquisition and processing

Participants underwent an imaging protocol including standard clinical sequences (FLAIR, DPT2- weighted) and a volumetric whole-brain 3D high-resolution T1-weighted sequence, performed with a 3 T Allegra MR imager (Siemens, Erlangen, Germany). Volumetric whole-brain T1-weighted images were obtained in the sagittal plane using MDEFT sequence. The FreeSurfer imaging analysis suite was used for cortical reconstruction of the whole brain.

2.5 Acoustic features

TSIA recordings were sampled to extract 10-sec samples from four interview items for each subject using Audacity. Automatic acoustic feature extraction was performed using the OpenSMILE API (v 2.0.2) in Python (v. 3.9.4). Univariate filtering with Spearman correlation was applied to select a set of features

associated with TSIA scores prior to Machine Learning modeling.

2.6 Machine Learning integration

Machine Learning algorithms were used to predict TSIA scores exploiting two different datasets: one made of integrated NLP, psychological, neuroimaging and demographic data and the other made of audio features extracted from OpenSMILE application, including demographics. Elastic net (EN) regression, principal components regression (PCR) and partial least squares regression (PLSR), and their results were used. The final models were selected by evaluating the R2 and root mean square error (RMSE) metrics of the models during cross-validation on both datasets

3. Results

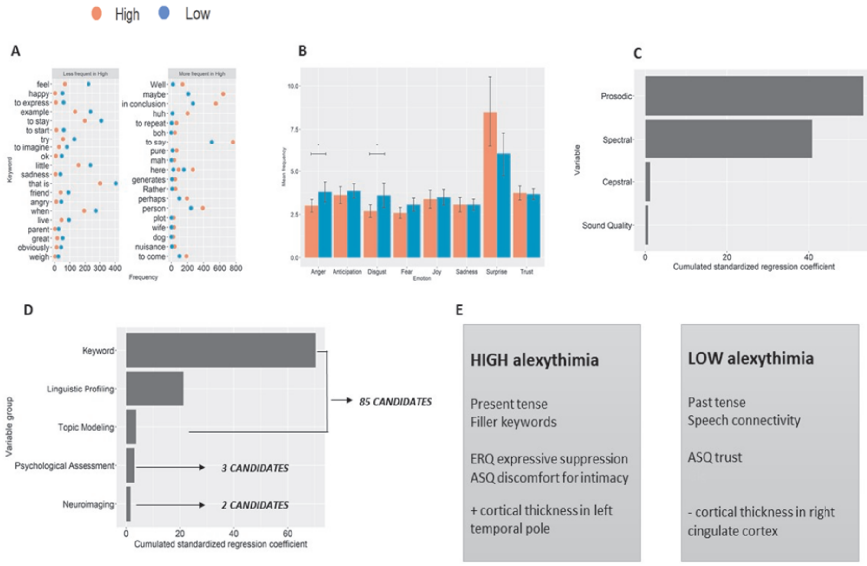


Figure 1. Sum of principal results.

Among NLP results, Semantic Analysis showed that subjects High Alexithymia used less emotional keywords and more filler words (Figure 1A). Unsupervised Latent Dirichlet Allocation (LDA) identified 20 topics, described by the top five words for frequency of use. Three topics were significantly different between High and Low Alexithymia subjects ($p \leq 0.05$; Topic 1: fashion, reason, end, minute, effect; Topic 2: day, home, week, sister, vacation; Topic 3: sense, parent, news, scene, patient). Exploring sentiment polarization, High Alexithymia subjects were characterized by a reduced frequency of emotion-related words, when compared to Low Alexithymia subjects, specifically for anger and disgust

(*at least p=0.025) (Figure 1B).
Machine Learning analyses on acoustic data revealed that Principal Components Regression (PCR) model with 5 components was the best fitting. On variable importance: sound quality, spectral, cepstral and prosodic domains were chosen by the model (Figure 1C). Machine Learning analyses on cross-disciplinary data revealed that Principal Components Regression (PCR) model with 3 components had the best performance over evaluation metrics. As for variable importance, 85 candidate features belonged to NLP, 3 to psychological measures, 2 to neuroimaging data, and 2 to demographic data (Figure 1D-E). The variable importance of neuroimaging data showed that the increased left temporal pole thickness predicted increased TSIA scores, while increased right isthmus cingulate thickness predicted decreased TSIA scores (Figure 1D-E).

4. Discussion

Given the language used to describe emotion indicates how one knows and experiences emotions (Lindquist, 2021), individuals with high alexithymic traits may be unable to accurately express their emotions verbally (Bird and Viding, 2014). The majority of studies on alexithymic subjects have focused on the occurrence of emotional words, but only a few have examined verbal encoding of emotion as a means of direct access to the symbolization of emotion. Specifically, people with high alexithymia use uncomplicated emotional language, and neuro-physiologically they show low sensitivity to the emotional qualities of speech (Goerlich et al., 2012).
In the present study, we combined psychological, linguistic and neuroimaging data to obtain a multidimensional analysis of the alexithymia. Study design and linguistic reference mapping were guided by several replicability and validity strategies (Straka et al., 2016). Feature selection, dimensionality reduction, and model independent testing techniques were used in the analysis pipeline to ensure reliable findings, in accordance with best data mining practices.
A Machine Learning approach was used to uncover specific characteristics in the linguistic profiles of individuals with high alexithymia. People with high alexithymia produced more language distortions, more frequently used the present tense of auxiliary verbs, used fewer possessive determiners, and had less interconnected speech patterns than people with low alexithymia.
Function words and conjunctions (e.g., articles, adverbs, and determiners) are automatically generated by language systems and serve to map relationships among the consciously generated meaning words (e.g., nouns or verbs) that carry the primary semantic information.
Accordingly, variations in the use of function words can provide sensitive indicators of emotional responses and predict the activation of automatic threat detection and response systems better than conventional self-report measures of

stress, depression, and anxiety (Mehl et al., 2017). High alexithymia speakers used fewer emotional keywords and filler words than other speakers; the latent semantic structure of their speech was dominated by body words, with a reduced semantic space for emotion. In fact, these participants were more neutral in the continuum of emotional valence and used fewer words associated with anger and disgust. Speaking about a specific emotion requires that emotion to be simulated, or partially re-experienced, which involves the same neural and bodily mechanisms recruited during that emotional experience. Accordingly, alexithymia can be defined as difficulty in elevating emotions from the sensorimotor to the symbolic level and separating internal physiological arousal from cognitive appraisal (Laricchiuta et al., 2014).

Regarding the acoustic analyses, prosodic and spectral features of language accurately predicted the levels of alexithymia. Variations in voice harmonicity, quality, and signal, three features of speech associated with arousal (Low et al., 2020), were negatively associated with alexithymia levels. Conversely, there was a positive association between alexithymia levels and microprosodic variations in the length of the fundamental frequency for harmonic sounds, which indicates irregularity in voice intensity and is usually associated with psychological distress (Low et al., 2020).

Using Machine Learning, this study combined linguistic data with demographic, psychological and neuroimaging data. We found that cortical thickness at the left temporal pole (positively) and at the isthmus of the right cingulate cortex (negatively) predicted alexithymia levels. Among individuals with high levels of alexithymia, structural and functional alterations are found in the frontotemporal structures, amygdala, insula, cingulate cortex, fusiform gyrus, parahippocampal gyrus, and cerebellum, known to play a role in emotional processing (Laricchiuta et al., 2015).

In particular, the amygdala and insula mediate emotional processing by identifying emotional stimuli (LeDoux, 2012). In the insula and the cingulate cortex, particularly the posterior subdivision and the right side, internal sensations are integrated with cognitive appraisal, resulting in affective awareness (Araujo et al., 2015). In cognitive appraisal, the frontal and temporal poles are involved for controlling emotions, and for linking interoceptive stimuli to their conceptual meaning. There has been a long-held assumption that alterations in the frontotemporal cortex contribute to impaired ability to express emotions symbolically and linguistically (Mendez, 2021).

Starting from the finding that the brain regions involved in semantics are also activated by emotional experience and perception (Lindquist, 2021), we propose that, in people with high levels of alexithymia, altered perception of visceral and somato-motor responses linked to emotional activation may be grounded in alterations in the isthmus of the right cingulate cortex. As a result of these

alterations, an individual may be unable to express emotions verbally and symbolically, due to changes to the left temporal pole. When interoceptive accuracy is impaired, somatic sensations may become dissociated from semantic knowledge, resulting in alexisomia (a difficulty in interpreting somatic symptoms), an important symptom in psychosomatic disorders (Mendez, 2021). Analyzing neuro-psycholinguistic data by using cutting-edge Machine Learning techniques offers a powerful approach for understanding the capacity of emotional experiences to be translated into verbal forms and identifying their brain structural correlates. As a result, the practice of using patterns of natural language as an indicator of implicit well-being may have a significant impact on health care and precision psychology.

References

Araujo H. F., Kaplan J., Damasio H. and Damasio A. (2015). Neural correlates of different self domains. *Brain and Behavior*, 5(12).

Bagby R. M., Taylor G. J., Parker J. D. and Dickens S. E. (2005). The Development of the Toronto Structured Interview for Alexithymia: Item Selection, Factor Structure, Reliability and Concurrent Validity. *Psychotherapy and Psychosomatics*, 75(1), 25-39.

Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S. and Matsuo A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774.

Bird G. and Viding E. (2014). The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews*, 47, 520-532.

Brunato D., Cimino A., Dell'Orletta F., Venturi G. and Montemagni S. (2020). Profiling-ud: a tool for linguistic profiling of texts. I*n Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 7145-7151)

Bucci W. (2021). Overview of the Referential Process: The Operation of Language Within and Between People. *Journal of Psycholinguistic Research*, 50(1), 3-15.

Eyben F., Wöllmer M. and Schuller B. (2010). Opensmile. *Proceedings of the International Conference on Multimedia* - MM 10.

Goerlich K. S., Aleman A. and Martens S. (2012). The Sound of Feelings: Electrophysiological Responses to Emotional Speech in Alexithymia. *PLoS ONE*, 7(5).

Laricchiuta D., Petrosini L., Picerni E., Cutuli D., Iorio M., Chiapponi C., . . . and Spalletta G. (2014). The embodied emotion in cerebellum: A neuroimaging study of alexithymia. *Brain Structure and Function*, 220(4), 2275-2287.

LeDoux J. E. (2021). What emotions might be like in other animals. *Current Biology*, 31(13).

Lindquist K. A. (2021). Language and Emotion: Introduction to the Special Issue. *Affective Science*, 2(2), 91-98.

Low D. M., Bentley K. H. and Ghosh S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96-116.

Mehl M. R., Raison C. L., Pace T. W., Arevalo J. M. and Cole S. W. (2017). Natural

language indicators of differential gene regulation in the human immune system. *Proceedings of the National Academy of Sciences*, 114(47), 12554-12559.
Mendez M. F. (2021). Frontotemporal Dementia: A Window to Alexithymia. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 33(2), 157-160.
Straka M., Hajic J. and Straková J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 4290-4297).

Framing the presence of tourism policies in the Calabrian public agenda during Covid-19 crisis: analysis of online news from local newspapers using 3-wave datasets

Sara Laurita[1], Rossella Nicoletti[1], Alfredo Fortunato[2], Carmelofrancesco Origlia[2]

[1]Wish Srls – laurita@wishinnovation.it
rossella.nicoletti@wishinnovation.it
[2]Contesti Srls – fortunato@contesti.info origlia@contesti.info

Abstract
The paper aims to study the online news published by the regional newspapers of Calabria in order to evaluate to what extent has influenced the definition of policies for recovery in the field of culture and tourism or has been influenced in turn. The use of automated web crawlers as a web content mining methodology and topic modelling techniques of Latent Dirichlet Allocation (LDA) lets us to determine the main topics that have marked, and are marking, the pandemic phase and the evolution of their incidence over the months. The in-depth analysis of the news items characterized by the prevalence of the topics of interest and the association of semantic similarity techniques makes it possible to assess their degree of similarity with the policy documents of the National Recovery and Resilience Plan. The analysis, focused on a territorial context, shows that the theme of tourism and culture is connoted by a limited relevance in the regional public debate, the evolution of which is affected by a driving effect rather than have an anticipatory one with respect to the definition and implementation of the NRRP.

Keywords: pandemic, online news, text mining, web content mining, LDA, tourism, culture, policy, recovery, resilience

Riassunto
Il lavoro ha come obiettivo lo studio delle notizie online pubblicate dalle testate regionali calabresi nel tentativo di comprendere se e come abbia influenzato la definizione delle policy per la ripresa in tema di cultura e turismo o ne sia stato influenzato a sua volta. L'uso di web crawler automatizzati come metodologia di web content mining e di topic modeling del Latent Dirichlet Allocation (LDA) permette di determinare i principali argomenti che hanno segnato, e stanno segnando, la fase pandemica e l'evoluzione della loro incidenza nel corso dei mesi. L'approfondimento sulle notizie caratterizzate dalla prevalenza dei temi di interesse e l'associazione di tecniche di similarità semantica consente di valutarne il grado di somiglianza con i documenti di policy del Piano

Nazionale di Ripresa e Resilienza. L'analisi, focalizzata su un contesto territoriale, evidenzia come il tema turismo e cultura si contraddistingua per una rilevanza limitata nell'ambito del dibattito pubblico regionale la cui evoluzione risente di un effetto traino più che avere un effetto anticipatorio rispetto alla definizione e all'implementazione del PNRR.

Parole chiave: pandemia, notizie online, text mining, web content mining, LDA, turismo, cultura, politiche, ripresa, resilienza

1. Introduzione

La pandemia da COVID-19 ha catalizzato per larghi tratti il dibattito pubblico nazionale e locale a partire dal febbraio 2020. Le notizie apparse sui siti web delle testate giornalistiche costituiscono una fonte rilevante, insieme ai social network, per capire forma ed evoluzione della discussione pubblica sugli aspetti epidemiologici così come su quelli maggiormente connessi alla vita economica e sociale. Obiettivo del paper è quello di analizzare le notizie online pubblicate dalle testate calabresi al fine di identificare i temi emergenti del dibattito regionale e di verificare se e in che misura questo abbia avuto un ruolo nella definizione delle policy per la ripresa post-pandemia in tema di cultura e turismo o ne sia stato influenzato a sua volta.

Numerosi sono gli studi che hanno indagato gli impatti della pandemia a vari livelli mediante tecniche di analisi testuale. La ricerca bibliografica è stata condotta mediante le seguenti keywords: "LDA", "Covid-19", "news online", "topic detection". I risultati ottenuti restituiscono un quadro certamente variegato di analisi nell'ambito del quale si possono riconoscere alcuni filoni ben distinti. Un set di studi è finalizzato a monitorare nel tempo l'evoluzione dei principali topic (Lygdman, 2022; Shah et al, 2021; Lyu et al., 2021; Ordun et al., 2020,), in qualche caso effettuando anche un'analisi del sentiment (Krawcyk et al., 2021; Abdulaziz et al.; 2021). Le sorgenti dati sono notizie (Lygdman, 2022; Krawcyk et al.; 2021), social media (Abdulaziz et al., 2021; Ordun et al., 2020), e documenti di ricerca (Carracedo et al.; 2021). Particolarmente rilevante ai fini del lavoro sono gli studi di Lygdman, 2022, e Lyu et al., 2021, che applicando la tecnica LDA ottengono un numero di topic che a loro volta vengono raggruppati in macro temi e studiati sia da un punto di vista temporale sia rispetto alla testata giornalistica. Lyu et al., 2021 evidenziano, poi, come il tema sui vaccini e la vaccinazione sia rimasto l'argomento più discusso sul periodo compreso tra marzo 2020 e gennaio 2021. Shah et al., 2021, e Krawcyk et al., 2021, ricercano argomenti e analizzano il sentiment rispettivamente di commenti su siti di rating medico e notizie online, al fine di determinare se e come cambiano le preoccupazioni delle persone in seguito alla pandemia. Shah et al., 2021, individuano 10 topic dominanti per ognuno dei tre archi temporali considerati nel biennio 2018-2020, dimostrando come l'epidemia determini mutamenti

significativi. Krawcyk et al., 2021 giungono alla conclusione che sia necessario veicolare ulteriormente l'informazione legata alle regole di distanziamento, per migliorare il livello di fiducia della popolazione.

2. Metodologia

Il set di dati è rappresentato da una collezione di news online di testate giornalistiche che hanno come ambito di riferimento il territorio della Calabria. Fra le testate giornalistiche operanti a livello regionale e regolarmente iscritte presso il Registro del Tribunale territoriale di appartenenza sono state selezionate quelle che soddisfacevano le seguenti condizioni:

- mostrare valori elevati di "popolarità" / "visibilità"[1];
- possedere un sito web sul quale era possibile programmare procedure di web scraping automatizzate.

L'analisi è stata condotta su sei testate online, ovvero Gazzetta del Sud, LaCNews24, Otto e Trenta, Quotidiano del Sud, Cosenza Channel e Reportage Online.

Per ogni notizia online sono state estratte le informazioni su: categoria di pertinenza, titolo, contenuto, data di pubblicazione, testata giornalistica di appartenenza, eventuali tag, sommario.

Il corpus e il relativo corredo informativo sono stati recuperati mediante procedure di web scraping in Python e sono stati sistematizzati all'interno di una base di dati di tipo SQL[2].

La categoria di pertinenza della notizia è stata utilizzata per una prima delimitazione del dominio di applicazione, in particolare sono state considerate le notizie inerenti alle categorie di "Politica", "News Calabria" e "Attualità"[3] con l'obiettivo di focalizzare l'attenzione sulle notizie che riportino informazioni su azioni, decisioni e programmazione pubblica. Si è così ottenuto un set di 15.544 notizie. Tale set di notizie è stato suddiviso in tre archi temporali:

- I wave coincidente con la fase pre-pandemia: dal 1 gennaio 2019 al 28 febbraio 2020;
- II wave riferita alla fase più acuta dell'emergenza sanitaria con l'applicazione di misure di contenimento più restrittive: dall'1 marzo 2020 al 25 aprile 2021;

1 Le metriche utilizzate sono quelle rese disponibili dal tool SimilarWeb (https://www.similarweb.com/).

2 Le procedure di web scraping automatizzate sono state progettate ad hoc per singola testata giornalistica e per singola categoria di pertinenza. Tali procedure implementate mediante Python (https://www.python.org/), ed in particolare con l'ausilio delle librerie BS4 (https://selenium-python.readthedocs.io/) e Selenium (https://selenium-python.readthedocs.io/)

3 Non sono state prese in considerazione le news categorizzate con i tag: cultura, viaggi, spettacolo, rubriche.

- III wave relativa alla fase di ripresa duratura della vita economica e sociale con l'accelerazione su vasta scala del piano di vaccinazione e l'invio del Piano Nazionale di Ripresa e Resilienza alla Commissione per l'approvazione finale: dal 26 aprile 2021 al 28 febbraio 2022.

La distribuzione delle news per testata e per wave non è omogenea. La presenza di testate a carattere locale o provinciale (es. Cosenza Channel) determina una maggiore copertura delle notizie relative ad alcune aree territoriali, pure molto rilevanti nel contesto regionale. Dapprima, il corpus di notizie è stato pre-trattato in coerenza con quanto emerso dalla letteratura mediante le seguenti operazioni: i) definizione di una di lista di stop-words; ii) conversione in lowercase, iii) eliminazione della punteggiatura in eccesso, lista di stop-words, punteggiatura, numeri; iv) indicizzazione del testo e nello specifico generazione di bigram; v) rappresentazione vettoriale del testo delle notizie mediante metrica del term-frequency (tf), ovvero matrice documento-termine (Allahyari et al., 2017, Vijayarani et al., 2015; Srividhya and Anitha, 2010). In seguito, è stata applicata la tecnica del Topic Modeling, nella sua declinazione del Latent Dirichlet Allocation (LDA), quale metodo probabilistico non supervisionato. Il numero di topic è stato impostato a 5 per le tre wave, a valle di analisi quantitative e qualitative. L'ipotesi è di individuare un set di argomenti nella prima wave da seguire nelle restanti, valutando se e come questi argomenti siano cambiati nel tempo. Inoltre i test eseguiti hanno dimostrato come il numero di topic pari a 5 sia quello che meglio si adatti alle tre wave tenendo conto delle misure definite da Griffiths and Steyvers, (2004), Cao et al., (2009), Arun et al., (2010), Deveaud et al. (2014).

Per ogni wave è stata così individuata la distribuzione dei topic emergenti. Con riferimento alla terza wave la tecnica è stata applicata una seconda volta considerando esclusivamente le notizie caratterizzate dalla prevalenza dei topic più direttamente connessi allo sviluppo economico e sociale identificati in precedenza, al fine di individuare i relativi sotto-topic. L'ipotesi formulata è che tali sotto-topic abbiano una probabilità maggiore di presentare riferimenti agli aspetti legati alla cultura e al turismo e che, quindi, possano costituire una proxy del nucleo semantico del dibattito pubblico regionale su questi temi.

Per valutare, poi, il grado di similarità con il testo del Piano Nazionale di Ripresa e Resilienza si è proceduto operando un'estrazione terminologica sulla parte del documento strategico riferita al turismo e alla cultura ricavando una bag of words di 1.074 termini ottenuta a valle delle medesime operazioni di pre-trattamento utilizzate per il corpus delle notizie. Questa è stata utilizzata per rappresentare in forma vettoriale il testo delle notizie utilizzando la metrica del term-frequency (tf) per la generazione della matrice documento-termine. Per determinare se e in che misura queste notizie fossero simili al testo del PNRR il confronto è stato condotto secondo due modalità: i) sulla frequenza dei termini della «bag of

words» fra il testo PNRR e il testo delle singole notizie secondo la misura similarità del coseno, ii) utilizzando l'approccio precedente a valle dell'applicazione del modello Latent Semantic Analysis (LSA). Con l'obiettivo di isolare le tendenze agli ambiti del turismo e della cultura, un ulteriore test è stato condotto applicando il topic modelling all'intero corpus e poi suddividendo nelle tre wave esclusivamente le notizie per le quali i topic più direttamente connessi allo sviluppo economico e sociale presentavano una prevalenza superiore al 50%. La similarità con il testo del PNRR è stata valutata con riferimento all'intero corpus di news ricadente nella terza wave.

3. Risultati

Nell'analisi del topic emergenti dalle *news online* relativamente ai tre periodi presi in esame occorre tenere conto di eventi importanti che, insieme alla pandemia, hanno caratterizzato il dibattito pubblico regionale. In particolare il 2019 è stato contraddistinto dalla campagna elettorale per l'elezione del nuovo presidente della regione, svoltesi a gennaio del 2020. Successivamente a causa della prematura scomparsa della presidente Santelli (ottobre 2020), sono state indette nuove elezioni regionali rinviate più volte a causa della situazione epidemiologica e poi svoltesi a ottobre 2021. Nel corso della medesima tornata elettorale, i cittadini di Cosenza, uno dei capoluoghi di provincia, sono stati chiamati alle urne anche per l'elezione del sindaco con il turno supplementare di ballottaggio. Inoltre, la tornata elettorale della primavera del 2022 interesserà la città capoluogo di regione Catanzaro.

I wave: Gennaio 2019 – febbraio 2020	**II wave: Marzo 2020 – marzo 2021**	**III wave: Aprile 2021 – febbraio 2022**
Topic 1(16.1%) Camera commercio; parco nazionale; beni culturali; marketing territoriale; turismo marketing	Topic 1 (21.9%) Campagna elettorale; partito democratico; raccolta differenziata; autorità portuale; forze politiche	Topic 1 (19.3%) Campagna elettorale; leader lega; commissione parlamentare; commissione antimafia; criminalità organizzata
Topic 2 (21%) Associazione vittime; forze ordine; vittime strada,; campagna elettorale; criminalità organizzata	Topic 2 (20.6%) Alta velocità; ponte stretto; corte conti; recovery plan; vittime strada_statale	Topic 2 (17.4%) Anti covid; terapie intensive; campagna vaccinale; casi covid; pari opportunità
Topic 3 (18.8%)	Topic 3 (16.7%)	Topic 3 (24,4%)

Tabella 1. Distribuzione topic e principali parole-chiavi del modello LDA per wave

I wave: Gennaio 2019 – febbraio 2020	**II wave: Marzo 2020 – marzo 2021**	**III wave: Aprile 2021 – febbraio 2022**
Corte conti; palazzo bruzi; dissesto finanziario; raccolta differenziata; piano riequilibrio	Isolamento domiciliare; domiciliare guariti; reparto rianimazione; rianimazione isolamento; reparto isolamento	Partito democratico; campagna elettorale; palazzo campanella; coalizione centrosinistra; segretario nazionale
Topic 4 (18.2%) Sanità calabrese; ministro grillo; commissario acta; servizio sanitario; piano rientro	Topic 4 (19%) Emergenza covid; sospensione attività; presente ordinanza; didattica distanza; ordinanza santelli	Topic 4 (19.2%) Alta velocità; ripresa resilienza; piano_nazionale ripresa; isolamento domiciliare; casi attivi
Topic 5(26.4%) Partito democratico; campagna elettorale; aiello callipo; candidato governatore; callipo santelli	Topic 5 (21.7%) Sanità calabrese; emergenza covid; commissario acta; sistema sanitario; sevizio sanitario	Topic 5 (19.7%) Sanità calabrese; open fascia; sistema sanitario; commissario acta; piano rientro

La prima wave si connota per la presenza di un topic relativo essenzialmente alla campagna elettorale regionale che è prevalente nel 26,4% delle notizie, seguito dal topic legato principalmente a viabilità e criminalità organizzata che ricorre come principale in poco più di una notizia su 5. Mostrano una prevalenza del 18,8% e del 18,2% rispettivamente un topic a carattere più miscellaneo che tocca diversi temi tra cui il dissesto finanziario dei Comuni, primo fra tutti quello di Cosenza, e un topic eminentemente votato alle problematiche del sistema sanitario regionale (livelli minimi di assistenza inadeguati, carenza di personale, ecc.). Relativamente meno rilevante con una prevalenza del 16,1% il topic riferito alle questioni dello sviluppo locale che in parte riguarda anche il settore cultura e turismo.

Nella seconda wave, conformemente alle aspettative, emerge con forza il ruolo della pandemia che riguarda almeno tre dei cinque topic rinvenuti con alcune declinazioni specifiche riferite rispettivamente a: i) ritardi strutturali del sistema sanitario regionale (21,7%), ii) mutamenti nella vita economica e sociali connessi alle restrizioni (19%), iii) contesto epidemiologico (16,7%). Completano il quadro un topic con una caratterizzazione politica ma abbastanza trasversale (21,9%) e un topic in cui compaiono per la prima volta gli aspetti connessi al

Piano di Ripresa e Resilienza insieme alle questioni legate alle infrastrutture ed ai trasporti (20,6%).
La terza wave, d'altra parte, è focalizzata sulle nuove elezioni regionali svoltesi il 4-5 ottobre 2021. Il topic che denota la prevalenza più elevata, pari al 24,4%, è interamente dedicato alla competizione elettorale, mentre ulteriori due topic trattano tali aspetti in connessione con la criminalità organizzata (19,3%) e l'emergenza sanitaria (17,4%). Inoltre, permane la presenza del topic connesso ai deficit del sistema sanitario regionale (19,7%) e del topic relativo al Piano di Ripresa e Resilienza insieme alle questioni legate alle infrastrutture ed ai trasporti inquadrate nella più ampia tematica dello sviluppo anche in connessione all'andamento epidemiologico (19,2%). Disaggregando quest'ultimo topic, emergono due sotto-topic che sembrano mettere in connessione il Piano di Ripresa e Resilienza con i ritardi strutturali del sistema sanitario (21,9%) e del sistema dei trasporti e della viabilità regionali (17,9%). Si registra, poi, un sotto-topic che guarda alle elezioni amministrative locali con particolare riferimento al Comune di Cosenza (21,5%) ed uno orientato all'ambiente e al rischio idrogeologico (18,5%). In ultimo si segnala la presenza di un sotto-topic a carattere più miscellaneo che collega le dimensioni dello sviluppo sostenibile, della transizione ecologica e della cultura (20,2%).
Dal confronto del testo delle singole notizie con la frequenza della *bag of words* di 1.074 termini estratti dalla sezione del PNRR focalizzata su turismo e cultura, la misura di similarità del coseno oscilla da un minimo dello 0,7% fino ad un massimo del 45%. Successivamente all'applicazione del modello LSA, il campo di variazione è compreso tra lo 0,7% e il 31,5%. L'approccio mediante LSA sembra riuscire a raffinare l'analisi. La notizia che riporta l'indice di similarità più elevato è relativa ad un incontro-conferenza promossa da Acli Terra Calabria sul pilastro cultura del PNRR quale occasione di sviluppo per il Sud.
L'ulteriore tentativo di analisi effettuato mediante applicazione del topic modelling all'intero corpus e successiva suddivisione nelle tre wave esclusivamente delle 4.667 notizie per le quali i topic più direttamente connessi allo sviluppo economico e sociale presentavano una prevalenza superiore al 50% non hanno modificato sostanzialmente i risultati ottenuti. Accanto ad alcuni temi caratterizzati da un'elevata stagionalità (es. incendi boschivi particolarmente rilevanti nella terza wave), sono emerse specifiche connotazioni della dimensione cultura e turismo (es. i parchi nazionali nella prima wave e il turismo delle radici nella terza) che pure vanno interpretate con cautela data la numerosità non elevata delle notizie prese in considerazione.

4. Considerazioni conclusive

L'analisi ha evidenziato come la pandemia abbia polarizzato il dibattito pubblico regionale, diventando catalizzatore dei ritardi che già contraddistinguevano la

Calabria e tema centrale delle importanti competizioni elettorali svoltesi nel periodo di riferimento. Inoltre, lo studio ha mostrato come il tema turismo e cultura si sia connotato per una rilevanza limitata nell'ambito delle *news online* e più che avere un effetto anticipatorio rispetto alla definizione e all'implementazione del PNRR sembra risentire di un effetto traino.
Gli sviluppi prossimi del lavoro mirano a filtrare le notizie sulla base di *keyword* legate al PNRR, al fine di eliminare possibili interferenze con articoli non di interesse e ad affinare la "scoperta" di topic con l'applicazione del Dynamic Topic Modeling al fine di "caratterizzare" ad un livello di dettaglio maggiore i contenuti memorizzati all'interno della base di dati.
Ulteriori obiettivi riguardano, poi, la qualificazione delle procedure di web scraping e web crawling nel tentativo di ampliare la platea delle testate giornalistiche considerate e di ragionare anche su altri contesti regionali con un'ottica di benchmark.

Bibliografia

Abdulaziz M., Alotaibi A., Alsolamy M., and Alabbas A. (2021). Topic based Sentiment Analysis for COVID-19 Tweets.Article in International Journal of Advanced Computer Science and Applications, vol. 12, pp.626-636.
Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E. D., Gutierrez J. B., and Kochut K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.
Arun R., Suresh V., Veni Madhavan C. E., and Murthy N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Pacific-Asia conference on knowledge discovery and data mining, Springer, Berlin, Heidelberg, pp. 391-402..
Cao J., Xia T., Li J., Zhang Y. and Tang S. (2009). A density-based method for adaptive lda model selection. Neurocomputing: An International Journal, vol.72 (7-9), pp. 1775–1781.
Carracedo P., Puertas R., and Marti L. (2021). Research lines on the impact of the COVID-19 pandemic on business. A text mining analysis. Journal of Business Research, vol.132, pp. 586-593.
Deveaud R., SanJuan E., and Bellot P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17 (1), pp. 61–84.
Griffiths T. L., and Steyvers M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), pp.5228-5235.
Krawczyk K., Chelkowski T., Laydon D. J., et all (2021). Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource. J Med Internet Res., 23(6):e28253.
Kuo H. Y., Chen S. Y., and Lai Y. T. (2021). Investigating COVID-19 News before and after the Soft Lockdown: An Example from Taiwan.Sustainability, vol.13(20), 11474.
Lygdman J, (2022). COVID-19 reporting in the news media in Finland: A Topic Modeling Approach.
Lyu J. C., Le Han E., and Luli G. K. (2021). COVID-19 vaccine–related discussion on

Twitter: topic modeling and sentiment analysis. Journal of medical Internet research, 23(6), e24435.
Ordun C., Purushotham S. and Raff E. (2020). Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. arXiv, arXiv:200503082.
Shah A.M.; Naqvi R.A. and Jeong O.-R (2021). Detecting Topic and Sentiment Trends in Physician Rating Websites: Analysis of Online Reviews Using 3-Wave Datasets. Int. J. Environ. Res. Public Health, vol 18.
Srividhya V., and Anitha R. (2010). Evaluating preprocessing techniques in text categorization. International journal of computer science and application, vol. 47(11), pp. 49-51.
Vijayarani S., Ilamathi M. J., and Nithya M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks,vol.5(1), pp. 7-16.

Poems and Songs.
What can Textual Data Analysis do?

Ludovic Lebart

CNRS (*R*) – ludovic@lebart.org

Abstract
Textual statistical techniques have proven their effectiveness in areas such as the processing of responses to open questions in sample surveys, automatic analysis of messages on the Internet, topic modelling. Applications of these techniques to creative works, such as plays or novels (typologies, seriations and chronological series, authorship attributions) are also possible, owing to the usual large sizes of these texts. But the most difficult endeavor is perhaps the application to poetry, or, worse, to songs, which are generally short texts with fanciful or incongruous lexical frequencies (figures of speech, metaphors, onomatopoeia, and, mainly for songs, choruses). We will show, based on the corpus of 194 songs of the French poet-singer Georges Brassens (1921-1981), that the tools of the statistical analysis of textual data, suitably adapted, can contribute to describing and understanding a poetic work, despite its disconcerting features. Owing to the specific status of lexical frequencies, binary tables (presence-absence) should imperatively replace frequency tables. Analyses should be performed on both lemmatized texts and original texts. Correspondence analysis supplemented by Clustering (mainly Additive Trees in this example) remains a robust description tool, provided that Bootstrap techniques are used at each step to validate the results.

Keywords: Poetry, Textual Data Analysis, Correspondence Analysis, Clustering, Bootstrap.

1. Introduction

The exploratory analyses of textual data have shown their usefulness in several fields of application: a) In the case of short, numerous and qualified texts such as the sets of responses to open questions in sample survey together with responses to closed questions and respondent characteristics (meta-data); b) In the case of large texts, such as corpuses of novels, political speeches, chronological textual series (evolution, discontinuities, possible attributions of authors). In these two main cases, the lexical frequencies are meaningful. But most sets of poetic texts do not fit into these frameworks. In the case of collections of comparable poems, such as Shakespeare's 154 sonnets, it has been shown that several statistical tools can detect topics (Topic modeling) already identified by experts in the past (see for example: Lebart, 2018; Lebart *et al*.,

2019). But the corpus that we propose to study here is a more complex and elusive material: the collection of 194 songs written and sung by the French musician-poet Georges Brassens (1921-1981). This author is unique in that he brings together three almost contradictory features: a) He was a nonconformist and has rubbed shoulders with anarchist movements, b) In 1967, he received the poetry prize from the very conservative *Académie Française*. c) He was at the origin of the sale of 30 million records.

2. Problems and challenges of poetic texts

We will begin by showing how the statistical analysis of this type of text constitutes a methodological challenge. Then we will describe the pre-processing of the corpus, the statistical procedures used, and will discuss the first results obtained. We will keep in mind the warning of Brunet (2004) during a remarkable statistical study of the poetic work of Arthur Rimbaud, a warning which applies also to the study of Brassens: *"Besides, the use of tools documentaries and statistics does not go without a certain naivety which gives its faith to the printed words, in their first innocence. But with Rimbaud, the words are often loaded. They are decoys, figureheads, and the reality they designate and hide eludes the most learned exegeses. When esotericism multiplies the traps, how to ensure the semantic constancy of the terms? "*. The studied works of Rimbaud included approximately 40,000 occurrences. The corpus of songs by Brassens used to exemplify the processing here has a comparable size: it contains approximately 52,000 occurrences. The poetic texts of Brassens are particularly rich in stylistic figures (litotes, metaphors, anaphors, euphemisms, allegories, etc.) which challenge the use of the word (being word-form or lemma) as a basic statistical unit. In that corpus, ribald, daring, even coarse slang words are sometimes inserted into the most subtle and tender poetry (Lamy, 2004, Poulanges and Tilleu, 2001). The poet brings up to date popular, outdated or slang expressions ("the poor man's coffee" for: sexual act, etc.). It often uses historical, obsolete terms (Rochard, 2009). In the case of songs that may include choruses or partial repetitions, versification constraints and rhymes, the lexical frequencies no longer have the statistical significance given to them in the lexical tables. Moreover, the texts are too short and the corpus too small to automatically extract new lexicographic units such as repeated segments, patterns, phrases.

The question we are asking may seem extremely narrow and technical in comparison with existing or potential works of literary and musical analyzes of the works of this musician-poet: "Do the typologies and visualizations obtained from the lexical profiles (approximately 1000 word-forms and as many lemmas) of the 194 songs, confronted with the available metadata, bring new information, relevant structural features or new materials likely to interest specialists?".

3. Preparation of texts and basic techniques

The 194 units of the corpus are described by as many groups of words (word-forms) and groups of lemmas which are complementary statistical units. These units are also described by the following metadata: a) Collection of albums (records) according to 14 categories, a variable which has a chronological component of publication (and often of composition/creation); b) The author according to 2 categories ("Brassens" [170 songs] or "Other", chosen and sung by Brassens [24 songs]); c) The dominant key (12 categories) of the musical score. The consideration of key is only a tentative experiment here. Note that the processing of musical scores through correspondence analysis has been studied, for example, in the pioneering works of Morando (1980), and Cocco (2014).

We will transform each song text, for both word-forms and lemmas, into an unweighted vocabulary, in other words, each element will only appear once within a given song (a few lines of Python easily allow this transformation of the text). The TreeTagger software (Schmid, 1994) has been used to obtain the lemmas. The lemmatized file has the advantage of reducing the diversity of inflections and therefore of allowing lower minimum frequency thresholds. It entails distances that are often more meaningful in terms of content whereas the file of word-forms retains the original diversity of forms, which is fundamental in the case of poetic texts. Thus, we will eventually obtain at each stage two different and complementary points of view.

Henceforth, the lexical tables (songs x words) will be tables of presence - absence. For this type of table, the correlation coefficient r between two songs coincides with the association coefficient Φ of Yule (1912). It is also linked to the χ^2 calculated on the contingency table (2 x 2) cross-tabulating the two texts described by n words, by the formula:

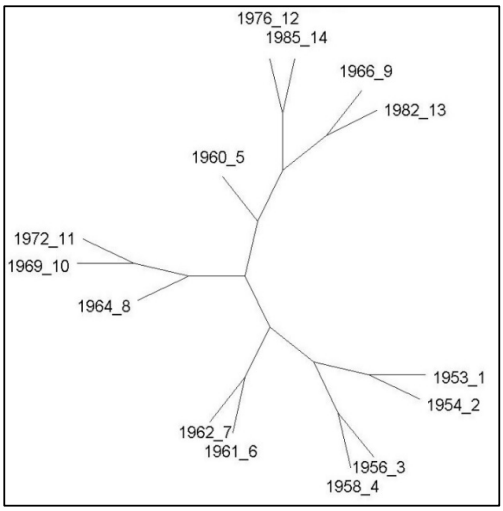
$$r^2 = \phi^2 = \frac{\chi^2}{n}$$

The techniques that can be used to describe and detect possible structures could be principal components analysis (licit on such binary coding), correspondence analysis (CA), clustering by additive tree (Huson and Briand, 2006), self-organized maps. The metadata defined above will act as active variables (grouping of songs according to records, for example), but above all as supplementary variables (*a posteriori* projection, with the indispensable Bootstrap validation). The software used, free of access, was DtmVic (www.dtmvic.com). Let us recall that in the context of binary coding, the eigenvalues do not have a straightforward interpretation in terms of information. Only the statistical validations (using the Bootstrap here) make it possible to judge the significance of the axes and the observed patterns.

4. Chronology and the sequence of records

The additive tree in figure 1 gives a summary of the links between the 14 main records (discs, albums) identified by their year of publication. The last two years (1982_13 and 1985_14) are posthumous. These albums may include songs composed previously, or assortments suggested by the publisher, or even, early works found lately and interpreted by others. The first four years (from 1953_1 to 1958_4) constitute the lower right part of the tree.

Figure 1. Additive tree of the 14 Brassens records (distances between records calculated from the presence-absence, within each song, of 943 word-forms).

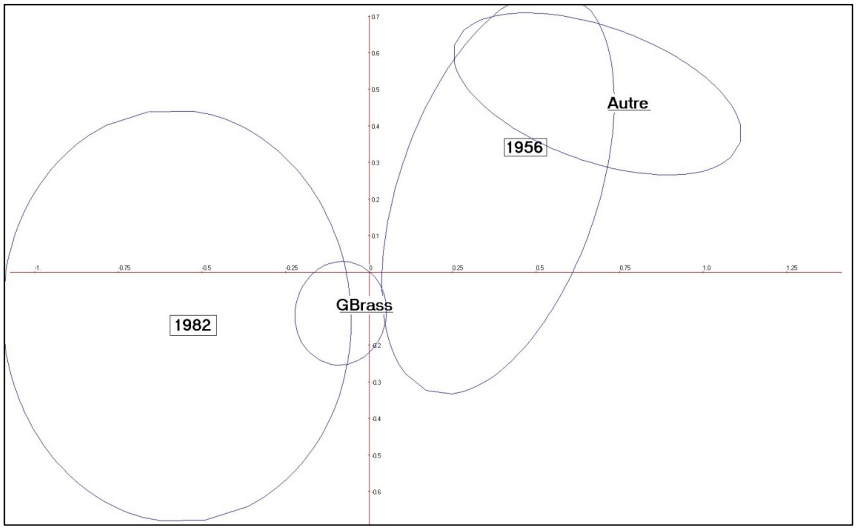


There is a certain compatibility between the lexicographic similarities described by this display and the chronology. The opposition between the first years and the following ones will be found in all the analyses, whether they use lemmas or word-forms.

5. Brassens and his selected extern poems

The 24 “external poems” set to music and sung by Brassens constitute a mini-anthology consistent with his personal values and his tastes. Figure 2 indeed shows the plan (1, 2) of the CA of the table crossing 194 songs and 901 lemmas (appearing at least four times).

Figure 2. Principal plane from the CA (194 songs, 901 lemmas). Bootstrap confidence areas for four supplementary categories. **GBrass** *(170 songs written by Brassens) and* **Autre** *(24 "external" poems set to music and sung by Brassens). Confidence areas for two records (***1982** *and* **1956***).*



The two "author" points and the two years chosen are projected *a posteriori* in the plane of the CA (194 songs, 901 lemmas). The Bootstrap replications are obtained by drawings with replacement within the set of 194 songs. The "external" authors (Hugo, Lamartine, Paul Fort, Musset, etc.) are closer to the early years (1956). In fact, one could see on the factorial maps containing the locations of the 901 lemmas (unpublishable in this publication format, but which will be available for consultation in the *www.dtmvic.com* website) that the vocabulary has hardened over time, the author calling himself at the end a "pornographer" (from record 5). At the same epoch, the censorship, active for the first records, was more tolerant over time.

6. Keys

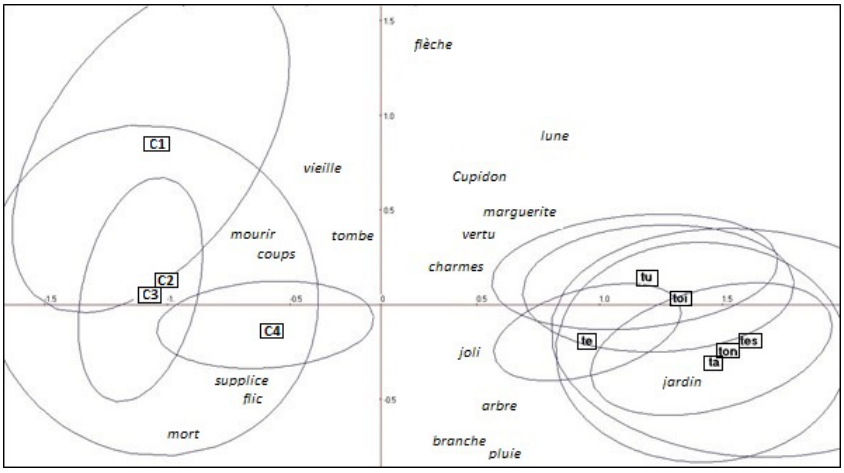
Brassens was a demanding composer-performer as well as a poet, and one could legitimately wonder if there was a link between the dominant keys and the characteristics of the songs represented by their lexical profiles. In the principal plane of the same CA (crossing 194 songs and 901 lemmas), the confidence areas of the 12 kept keys are indistinguishable except for G minor (Gm) on the left hand part of the display. However, this Gm exception only concerns five songs, three of which were posthumously performed by other singers. We will conclude,

at this current state of research, that there are no significant links between the chosen key and the lexical profiles of the songs.

7. Confrontations between lemmas and word-forms

This phase is the most essential part when dealing with a poetic text. We will work on the 170 texts written by the singer himself. For most of the CA-type analyzes bearing on the 194 songs, or on the 170 songs of which Brassens is the sole author, a first dimension (horizontal on all the displays) dominates, whether it concerns lemmas or word-forms: It opposes ancient texts (four to six first records) to more recent ones. Most of the ancient texts, like the external poems, have (often) a vocabulary that can be described as classic, even gallant, courtly or precious. The most recent have more cruder vocabulary. Figure 3, which concerns the 170 songs written and composed by Brassens, contains only a small excerpt of the set of the 872 active form words.

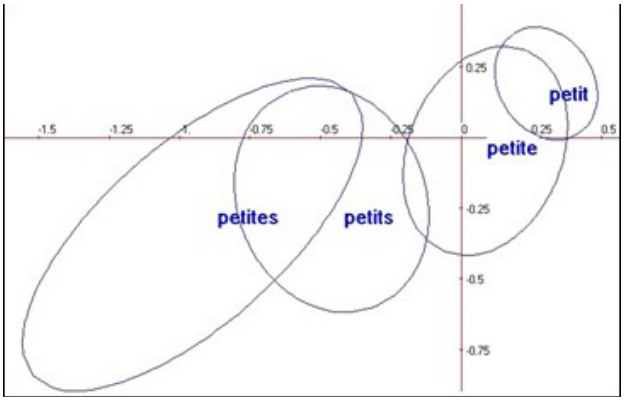
*Figure 3. Word-forms (Brassens alone) (170 songs x 872 forms). In the foreground of the CA, opposition between the forms of familiarity for "you" in French (***te, tu, toi, ton, ta, tes,** *on the right) linked to a refined poetry, and the presence of more bawdy vocabulary (on the left)* **C1, C2, C3, C4** *(cuckold, butts, ass, asses). Some Bootstrap confidence zones are drawn around the boxed item. The plan is also illustrated with some word-forms.*



It shows that the familiarity (six forms of *you* concerned on the right) means above all intimacy, gentleness for Brassens (except for one song). Here, the word-forms reinforce the interpretation of this axis which is similar to the one obtained with the lemmas. In addition to the previous word-forms, we also find among the 20 rightmost points on this axis the words (*arrow, daisy, moon, Cupid, virtue,*

charms, garden, rain, beautiful, tree, branch). On the left-hand part of the display, the four confidence areas concern less polite words are not commented here.

*Figure 4. Bootstrap confidence zones of four inflections of the adjective "*petit*" (small) in the CA first plane of the presence-absence array (194 songs x 872 word-forms).*



The word-forms play a role of magnifying glass for the interpretation, but they reveal also words that the lemmatization makes disappear, as shown in figure 4. In a CA performed on lemmas, the four inflections of the lemma "petit" in Figure 4 would be replaced by the unique lemma "petit", located in a central position which, in a way, neutralizes its contribution to the axis. These different inflections are markedly linked to the dimension which opposes, as we have seen, the intimate and the delicate (singular) to the impersonal and the bawdy (plural).

8. Conclusions

Despite the limited number of presented displays (necessarily small sized), one can guess that multivariate descriptive processing of poetic texts provides an original point of view on these texts and, above all, new material for analysis. Even from these first attempts of analysis, one could detect a general tendency, inextricably linked to age, career, personal evolution, perhaps to the increasing notoriety of the poet and probably to the growing permissiveness of the time. Each time, the use of forms amplifies, illustrates, and nuances the results obtained from the lemmas. The non-published graphs of the large additive trees representing the links between the 194 songs, the displays of principal planes from CA involving about 900 word-forms and as many lemmas together with the songs and the records, constitute also a set of working documents full of potential, difficult to publish on paper, but fascinating to consult for specialists

or amateurs. We will risk parodying the adage "Garbage in, garbage out" (adage in fact very poorly adapted to data analysis techniques that do have a filtering function) by transforming it into "Poetry in, poetry out". Indeed, these new documents describing the intricate links and patterns between hundreds of words carefully chosen by the great troubadour Brassens are themselves a source of poetic emotions.

References

Brunet E. (2004). Statistiques Rimbaldiennes, SI@T, *Les littératures de l'Europe unie*, Cesenatico, Italie, 88-113, hal-01362731.

Cocco, C (2014). *Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation.* Thèse, Université de Lausanne, Faculté des Lettres, Switzerland.

Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

Morando B. (1980). L'analyse statistique des partitions de musique, *Les cahiers de l'analyse des données*, tome 5, no 2 (1980), 213-228, http://www.numdam.org/item?id=CAD_1980__5_2_213_0

Lamy J.C. (2004). *Brassens, le mécréant de Dieu.* Albin Michel, Paris.

Lebart L. (2018). Looking for Topics, a brief review. In: *Text Analytics, Advances and Challenges.* Iezzi D. F., Mayaffre D., Misuraca M. (Eds), Springer, Cham, Switzerland, 215-224.

Lebart L., Pincemin B., and Poudat C. (2019). *Analyse des Données Textuelles*. P.U.Q. Québec, Canada.

Poulanges A. and Tilleu A. (2001). *Les manuscrits de Brassens*. Textuel, Paris.

Rochard L. (2009). *Les mots de Brassens*, Edition du Cherche Midi, Paris.

Yule, G. U. (1912). On the methods of measuring the association between two attributes, *Journal of the Royal Statistical Society*, 75, 579-642.

Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Caractériser les logiques disciplinaires et institutionnelles dans les rapports d'autoévaluation

Virginie Lethier[1], Emilie Née[2], Hugo Dumoulin[3]

[1]Université de Bourgogne Franche-Comté – virginie.lethier@univ-fcomte.fr
[2]Université Paris Est Créteil – emilie.nee@u-pec.fr
[3]Université Paris Nanterre – hdumoulin@parisnanterre.fr

Abstract
This paper studies self-evaluation reports of research laboratories in France. We aim to test the hypothesis of discipline and institutional position (status and size) as factors of discursive variation. For this purpose, a principal component analysis (PCA) is deployed on a corpus of 38 reports, articulating 60 descriptors of morphosyntactic order.

Keywords: Discours analysis; academic discourse; reports; genre; parts of speech

Résumé
Cette contribution a pour objet d'étude les rapports d'autoévaluation des laboratoires de recherche en France. Il s'agit de tester l'hypothèse de la discipline et de la position institutionnelle (statut et taille) comme facteurs de variation discursive. Pour cela, est déployée sur un corpus de 38 rapports une Analyse en Composantes Principales (ACP), articulant 60 descripteurs d'ordre morphosyntaxiques.

Mots clés : Analyse du discours ; discours académique ; rapport ; genre ; morphosyntaxe

1. Introduction

C'est à l'une des formes de la littérature grise académique qui témoignent des mutations de la sphère de l'activité scientifique (Sargeac, 2018) que s'intéresse notre contribution : le rapport d'autoévaluation des laboratoires de recherche universitaires français. Instrument d'évaluation externe de l'activité scientifique, le rapport d'autoévaluation correspond à un écrit standardisé dont la production est fortement encadrée par le Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur (Hcérès). Cette standardisation visée par l'Hcérès est-elle néanmoins effective ?

Il paraît légitime de supposer que le rapport d'autoévaluation, parce qu'il est un écrit hybride entre discours scientifique et discours managérial, est susceptible

d'être soumis à des variations (énonciatives, micro-textuelles ou stylistiques, argumentatives) imputables aux logiques disciplinaires et institutionnelles qui traversent la sphère d'activité scientifique. Plus précisément, la discipline dont relève l'unité de recherche qui produit le rapport, sa position institutionnelle et sa taille nous semblent constituer des facteurs de variation de premier plan.
Après avoir présenté brièvement les caractéristiques du rapport d'autoévaluation et les problématiques qu'il soulève, nous rendrons compte de nos choix méthodologiques destinés à émerger, dans une perspective exploratoire, les principes et les lieux textuels de variations de ce genre. Dans un dernier mouvement, nous présenterons les pistes ouvertes par une analyse en Composantes Principales.

2. Contexte de la recherche

2.1. Un écrit hybride, au caractère obligatoire et standardisé

Témoignant du caractère contractualisé de la recherche scientifique en France, le rapport d'autoévaluation correspond à un écrit de commande produit dans le cadre du processus d'évaluation de l'activité scientifique organisé par l'Hcérès. L'agence évaluatrice contraint fortement la production de ces rapports, dont la comparaison serait facilitée par leur standardisation. En amont, le rapport est un formulaire structuré et encadré par des référentiels ; en aval, il est un imprimé à remplir (Barats, Née, 2020). L'organisation en parties et sous-parties, les intitulés des (sous-)titres, les séquences textuelles qui les composent et la façon dont celles-ci s'articulent sont prédéfinies. Loin de n'affecter que le niveau macrotextuel, la trame imposée contraint tous les versants de la textualité : elle fixe les thématiques à traiter, les outils cognitifs à manipuler, le lexique et les formules à imprimer. Peu propices à la réflexivité cette logique formulaire et le format bref imposé (Rastier, 2013) rentreraient en tension avec l'objectif affiché du rapport (s'auto-évaluer). Par ailleurs, si le rapport affiche une visée informative, puisqu'il s'agit de « rendre compte » d'une activité scientifique, il est tout autant caractérisé par une visée argumentative : le rapport est un espace d'autolégitimation et de valorisation de l'unité de recherche dans un contexte concurrentiel que l'exercice participe à alimenter.
Le rapport d'autoévaluation se situe à l'interface de deux types de sphères d'activité, et par suite, de deux types de discours : le discours scientifique d'une part, le discours managérial de l'autre. Appelant à s'acculturer à un lexique, à un « discours (et une rhétorique) de projet », mêlant technicité et idéalité (Sargeac, 2018 : 64), cet écrit est à rédiger sur la base d'outils cognitifs issus de l'entreprise comme la matrice « SWOT ». On notera d'ailleurs que ce sont les codes sémiotiques de la brochure d'entreprise qui cadrent la mise en page de certains rapports d'autoévaluation. Pour autant, en raison de l'objet même du rapport (l'activité scientifique), de la culture socio-professionnelle des scripteurs

et de ses modalités concrètes de production (recyclage et montage de documents, écriture collective), le rapport peut être considéré comme l'un des genres du discours scientifique. Le rapport d'autoévaluation figure ainsi un entre-deux discursif, dans l'espace textuel duquel l'acculturation et/ou l'adhésion aux formats et aux formes du discours managérial sont susceptibles de fluctuer.

2.2. Hypothèses et questions de recherche

Une abondante littérature s'est attachée à problématiser et décrire les caractéristiques des pratiques diversifiées et des genres pluriels (articles, communications orales) que recouvre la catégorie de « discours scientifique ». Différentes études empiriques (Grossmann, 2010 ; Rinck 2010 ; Fløttum et Vold, 2010, Poudat et Follette 2012) ont ainsi démontré l'existence de logiques disciplinaires ou sectorielles (grands ensembles de disciplines) à l'origine d'importantes variations, affectant notamment les marques de l'énonciation. En nous inspirant de cet état de l'art, nous émettons à notre tour la première hypothèse selon laquelle le rapport d'autoévaluation, en tant que genre, est soumis à un principe de variation d'ordre disciplinaire : la discipline dont relèverait l'unité de recherche productrice du rapport déterminerait les dimensions plurielles de la textualité.

Par ailleurs, en France, les unités de recherche ne sont nullement des institutions aux caractéristiques intrinsèques homogènes. Elles peuvent correspondre statutairement à deux types de structure : les équipes d'accueil (EA) d'une part, les unités mixtes de recherche (UMR), labellisées par le Centre National de la Recherche Scientifique (CNRS). Qu'elles soient des EA ou des UMR, les unités de recherche peuvent être mono- ou inter-discipinaire. Leur taille, c'est-à-dire leur dotation en personnels statutaires, varie enfin sensiblement. Ces variables constituent, selon nous, des sources potentielles de variation des caractéristiques des rapports d'autoévaluation.

Parmi l'ensemble de ces variables externes (discipline, statut, taille, consistance mono- ou inter-disciplinaire), lesquelles contribuent-elles prioritairement à expliquer les variations observables dans les rapports d'autoévaluation ?

3. Méthodologie

3.1. Corpus

Notre corpus contient 38 rapports d'autoévaluation – soit 921989 occurrences. Homogène en termes de genre, ce corpus présente plusieurs facteurs de variation. Il intègre ainsi 12 rapports d'unités UMR (labellisation signalée dans le nom du CODE renvoyant à l'UR), contre 26 rapports d'équipes d'accueil. Par ailleurs, ces unités se subdivisent en deux types : monodisciplinaire ou interdisciplinaire. Notre corpus compte ainsi 24 rapports d'unités monodisciplinaires, couvrant les principales disciplines des Sciences humaines :

Discipline	Codes
Droit (n=7)	DROIT1, DROIT2, DROIT3, DROIT4, DROIT5, DROIT6, DROIT7UMR
Langues (n=3)	LANG1, LANG2, LANG3
Histoire (n=2)	HIST1, HIST2
Littérature (n=2)	LITT1, LITT2
Archéologie (n=1)	ARCH1UMR
Economie (n=1)	ECO1UMR
Gestion (n=1)	GESTION
Info-com (n=1)	INFOCOM
Linguistique (n=1)	LINGUMR
Sciences po. (n=1)	POL1UMR
Psychologie (n=1)	PSYCHO
Thermique (n=1)	THERM1

Et intègre 14 rapports d'unités interdisciplinaires, articulant les disciplines listées ci-dessous :

Codes	**Disciplines**
INTERA	Physique, mathématique, mécanique, thermique
INTERarUMR	Archéologie, ethnologie
INTER1UMR	Histoire, sociologie
INTER2	Histoire, philosophie
INTER3	Sociologie, philosophie
INTER4UMR	Sciences politiques, sociologie
INTER5	Sociologie, sciences du sport, Médecine
INTER6UMR	Ethnologie, sociologie
INTER7UMR	Géographie, sociologie
INTER8UMR	Architecture, urbanisme, sociologie, anthropologie, géographie
INTER9UMR	Histoire, anthropologie
INTERpsy	Psychologie, psychanalyse
INTERcog	Psychologie, Sciences du sport
INTERlang	Etudes slaves, germaniques, histoire du droit

Enfin, sur la base de leurs ressources humaines fixes, 22 des unités de recherche considérées correspondent à des unités de petite taille[1] (en noir ci-dessus et dans le graphique), 10 à des unités de taille moyenne (en bleu), 6 à des unités importantes (en vert).

3.2. Jeu de données

Notre jeu de données articule 60 descripteurs d'ordre morphosyntaxiques, niveau moins sensible aux variations disciplinaires que le vocabulaire. 32 de ces

[1] La masse salariale annuelle M sépare les laboratoires en groupes relativement distincts selon leur petite, moyenne ou grande taille (M < 2 250 000€ ; 2 250 000€ < M < 3 250 000€ ; M > 3 350 000 €).

descripteurs sont relatifs à la ponctuation et aux catégories grammaticales, telles que traditionnellement reconnues et étiquetées automatiquement via TreeTagger[2], appelé par TXM (Heiden et al., 2010), logiciel employé pour segmenter et indexer notre corpus. Ces premières indications ont été complétées par un jeu de 22 descripteurs issus de l'analyseur Cordial[3], qui renseignent notamment le type de propositions et le type de phrase (verbale ou non) du corpus, et différentes caractéristiques formelles telles que le nombre de lettres par mots. Certains des descripteurs ont été sélectionnés en vue d'affiner la lisibilité des catégories grammaticales traditionnelles, à l'instar des cinq sous-types d'adverbes (lieu, temps, manière, doute, négation) que l'analyseur se propose de repérer, ou du caractère abstrait/concret des noms communs. Six descripteurs spécifiques aux caractéristiques du genre du rapport d'évaluation et aux objectifs de notre recherche ont enfin été intégré : ils se rapportent à l'emploi des principaux pronoms personnels utilisés dans l'article (*il impersonnel*[4], *on, nous*). Concernant le pronom nous, sa place syntaxique a été renseignée (*sujet, objet*) au terme d'une procédure d'annotation manuelle.

4. De l'analyse en composantes principales

En vue de mettre au jour les interrelations entre ces descripteurs et de repérer, parmi les rapports d'autoévaluation, des groupes homogènes ou atypiques, nous nous sommes appuyés sur une Analyse en Composantes Principales (ACP). Cette méthode factorielle présente l'intérêt de réduire la complexité de la matrice initiale en un faisceau réduit de variables artificielles.

4.1. Vue sur l'ACP

L'axe 1, qui représente 20.06% de la variance, marque une opposition principalement relative au degré de complexité des phrases. Dans son versant gauche, se concentrent les propositions indépendantes et les indices d'un style à consistance nominale (*noms communs, adjectifs, noms concrets*), tourné vers la précision et l'énumération (*parenthèses et deux points*) et faisant la place belle aux données chiffrées (*numéral, chiffres, adjectifs cardinaux*). À droite de l'axe 1 se donne à lire le profil d'une écriture plus complexe (*propositions subordonnées et relatives, pronoms relatifs et démonstratifs, virgule*), à la consistance plutôt verbale (*phrases verbales, adverbe, verbes*) recourant à l'ensemble des pronoms personnels.

[2] Institute for Computational Linguistics, University of Stuttgart.

[3] Synapse Développement.

[4] Seuls les cas du « il » dans une construction impersonnelle ont été comptabilisés grâce à une recherche exploitant les possibilités du Corpus Query Language sur corpus étiquetés, dans TXM

L'axe 2, qui extrait 9.04 % du nuage de points, contribue à singulariser dans sa partie haute une énonciation subjectivisée et embrayée, centrée sur le « nous » (PRO1PP, verbe1PP, PRO1PP_sujet). Si l'axe est négativement corrélé aux verbes conjugués à la troisième personne du singulier, ses points extrêmes correspondent à deux marques de ponctuation (*guillemets, tirets*) dont la pluralité des fonctions en discours entrave toute interprétative sans retour au texte.

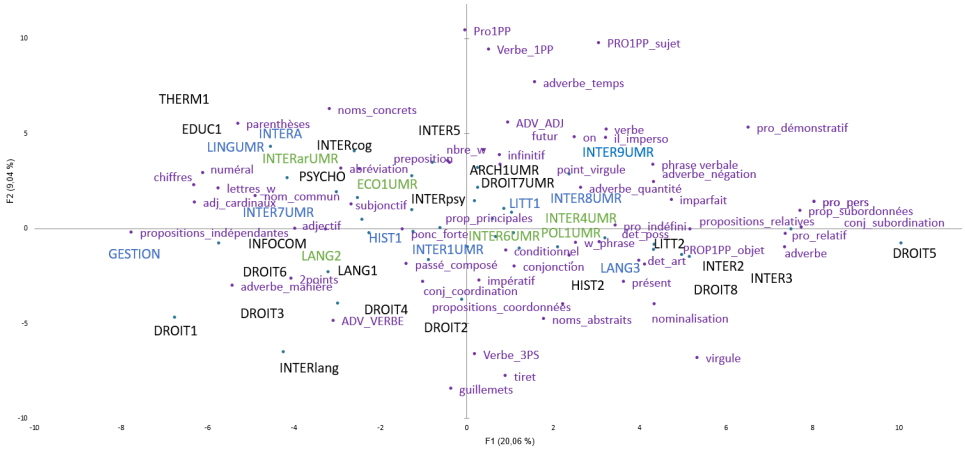


Figure 1. Visualisation des individus et des variables actives

4.2. Eléments d'interprétation

Si l'on considère la position des unités de recherche sur ce graphique, on repérera que le premier facteur semble prioritairement refléter des logiques sectorielles, affinées (ou contrariées) sur le deuxième facteur par le statut du laboratoire et sa taille. Le facteur 2 tend en effet à attirer, dans sa zone supérieure, les UMR et/ou de taille moyenne et forte, pour reléguer la plupart des équipes d'accueil dans sa zone basse. Le caractère mono-ou interdisciplinaire n'aurait, pour sa part, qu'une très faible valeur explicative. Sous l'effet conjoint des deux premiers facteurs, quatre regroupements notables sont à repérer.

Dans le quart supérieur ouest du graphique, s'entremêlent les sciences physiques (THERM1, INTERA) et certaines sciences humaines dites formelles et/ou outillées : l'économie (ECO1UMR), les sciences de la cognition et de l'expérimentation (PSYCHO, INTERcog, INTERPSY), la linguistique (LINGUMR). C'est dans cette zone que s'inscrivent l'UMR d'urbanisme (INTER7UMR), l'unité interdisciplinaire tissant archéologie et ethnologie (INTERarUMR), ainsi que les sciences de l'éducation (EDUC1).

Le quart supérieur est du graphique tend, pour sa part, à concentrer des unités UMR ou de taille moyenne tournées vers les sciences sociales, à l'instar des unités interdisciplinaires INTER9UMR (histoire, anthropologie) et

INTER4UMR (sciences politiques, sociologie). Il semblerait que l'inscription de l'unité liée au Droit dans cette zone soit principalement liée à son statut d'UMR. C'est en effet dans le quart inférieur ouest du graphique que se concentrent la plupart des autres unités de droit et de langues (LANG1, LANG2), la gestion et les sciences de l'information-communication, ainsi qu'un laboratoire d'histoire de taille moyenne.
Dans le quart inférieur est du graphique, on retrouvera principalement de petites équipes d'accueil, s'inscrivant dans des disciplines principalement déjà représentées dans la partie ouest de la zone basse du graphique (le droit, les langues, l'histoire), ainsi que deux unités interdisciplinaires partageant de prendre appui sur la philosophie (INTER2 ; INTER3).

5. Synthèse

Dans cette contribution, une analyse en composantes principales a été menée en vue de mieux cerner la façon dont des variables extra-discursives se distinguaient et s'articulaient pour expliquer les différences d'écriture des rapports d'autoévaluation. Cette première classe d'analyse, qui doit bien entendu être corroborée par une méthode de classification hiérarchique (CAH), confirme que les rapports d'autoévaluation diffèrent en fonction des disciplines et du statut institutionnel des unités de recherche, et plus discrètement, de sa taille.
D'autre part, nous attendions de cette analyse exploratoire qu'elle oriente nos regards d'analystes du discours vers les lieux morphosyntaxiques matérialisant ces variations. En l'occurrence, la complexité syntaxique des phrases, les marques de l'énonciation, la séquentialité (description prenant la forme de listes ou de développements plus élaborés), l'usage ou l'évitement des embrayeurs, la temporalité (verbes et adverbes), enfin le lexique abstrait et les nominalisations sont autant d'observatoires à prendre en considération pour des recherches ultérieures.

Bibliographie

Barats C., Née É. (2020). « Structurer, encadrer et lisser l'évaluation : les écrits de l'AERES/HCERES ». Communication et langage 203 : 5-25. DOI 10.3917/comla1.203.0005.
FlØttum K., Vold E.T. (2010). « L'éthos auto-attribué d'auteurs-doctorants dans le discours scientifique ». Lidil 41 : 41-58. DOI : 10.4000/lidil.3006.
Grossmann F. (2010). « L'Auteur scientifique. Des rhétoriques aux épistémologies ». Revue d'anthropologie des connaissances, vol. 4/3 : 410-426. DOI 10.3917/rac.011.0410.
Heiden S., Magué J.-P., Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010, Rome, pp. 1021-1032.

Née E., Oger C., Sitri F. (2017). « Le rapport : opérativité d'un genre hétérogène ». Mots 114 : 9-24. DOI 10.4000/mots.22752
Rastier F. (2013). Apprendre pour transmettre. L'éducation contre l'idéologie managériale. PUF.
Sargeac M. (2018). Le paradoxe de l'isomorphisme dans la concurrence ? Une analyse des discours par et sur les établissements de l'enseignement supérieur français. Mémoire de master 2, Paris, EHESS/ENS.
Poudat C., Follette P. (2012). "Corpora and academic writing: A contrastive analysis of research articles in biology and linguistics". In Boulton, A., Rowley-Jolivet, E., Carter-Thomas, S., Corpus-Informed Research and Learning in ESP: Issues and Applications. John Benjamins.
Tutin A. (2010). « Dans cet article, nous souhaitons montrer que… Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines ». Lidil 41 : 15-40. DOI : 10.4000/lidil.3040.

Automatic Anonymization of Italian Legal Textual Documents using Deep Learning

Daniele Licari, Maria Francesca Romano, Giovanni Comandé

Scuola Superiore Sant'Anna – daniele.licari@santannapisa.it
mariafrancesca.romano@santannapisa.it
giovanni.comande@santannapisa.it

Abstract
The dissemination of judicial decisions not only provides a valuable source of decision support for judges and legal practitioners but also strengthens public confidence in the judicial system. However, the nature of the data raises privacy concerns as the documents include personal and, often, sensitive data such as health, financial, religious beliefs, sexual orientation, and so on. In recent years, especially since the introduction of GDPR, the international scientific community has paid much attention to the issue of privacy and automatic anonymization tools, but unfortunately, nothing has been done in the Italian legal context.
In this paper, we present a first solution aimed at automatic anonymization of the Italian National Jurisprudential Archive (Archivio Giurisprudenziale Nazionale) domain based on pre-trained Transformers embeddings (Clark et al., 2020, Devlin et al., 2019) and spaCy's transition-based parsing for entity recognition (Honnibal and Montani, 2017). It achieves more than 94.7% recall (>99% for Person and ID entities) and supports several anonymization methods that can be applied to the text depending on the purpose of anonymization.

Keywords: anonymization, pseudonymization, GDPR, Italian court documents, NER, Deep Learning.

1. Introduction

Databases built on published judicial decisions are critically important for developing decision support tools and transparent administration. However, these openly published data can contain much sensitive personal information that should not be published. Automatic text anonymization is a solution to remove all sensitive information from documents, but it is not easy due to the unstructured form of textual data. Many current research projects have been published on this topic. Automatic or semi-automatic pseudonymization tools are developed for Finnish court judgments (Oksanen et al, 2019), for Uruguay

National Jurisprudence DataBase (Garat and Wonsever, 2022), for German legal court decisions (Glaser et al, 2021), for French legal texts (Barriere and Fouret, 2019), for documents written in Portuguese (Mamede et al. 2016), in Polish (Oleksy et al, 2021) or for English texts (Kleinberg et al., 2017).
The Italian National Jurisprudential Archive (Archivio Giurisprudenziale Nazionale, pst.giustizia.it) contains a selection of judicial decisions in their original form. This service, provided by the Ministry of Justice, was created to be shared only between judges. The service aimed to save judges time in their decision-making process and allow for the monitoring of compliance.
In recent years, the archive has been accessible to any Italian resident in possession of a digital identity (Sistema Pubblico di Identità Digitale, SPID) raising privacy issues. Current practice requires that documents are anonymized manually before being uploaded into the archive. This is very time-consuming and, often, documents are processed incorrectly; for example, being uploaded without removing personal information.

2. Related Work

In this paper, we argue and demonstrate that the current approach of manual anonymization is an insufficient method to both preserve privacy and cope with the massive amount of legal data. We propose an automated approach based on Deep Learning to extract and anonymize personal data from unstructured text. It can lead to significant savings in manual labor and minimize the time required to create a privacy-preserving (GDPR-compliant) publishing system of legal data.
The main information to be anonymized may concern the parties; testimony; important companies, identifiers, dates, or places. We have grouped into five categories of entities as shown in Table 1.

Table 1. Entity name and description of the 5 entities to be detected

Entity	Description
Company	Companies, agencies, institutions, organization
Location	Place of birth, marriage, death, residence, addresses, countries, cities, states
Date	Date of birth, marriage, death, residence, date of events.
Person	Names of parties, lawyers, judges, family, experts, witness
ID	Tax Codes, number plates, VAT numbers

3. Dataset

The civil law cases used to train and evaluate our Named Entity Recognition (NER) models come from the Civil Law Database of Genoa's Trial Court. It contains metadata and PDF documents about decrees, ordinances, and judgments from 2008 to 2019. The database includes around 24.000 judgments across the different subjects. Given the long experience of our research team, we have

focused our analysis on family (n=6000) and personal injury damage (n=59) judgments after 2014 since from that date the digitalization of the court decision documents was completely introduced in the Italian Legal Courts. Our NER models were trained and tested on a sample containing all personal injury sentences and an equal number of family sentences selected stratified on text length (number of words). Finally, the selected PDF documents were converted into texts using the open-source Tika python library and, then, imported to the annotation tool Doccano.

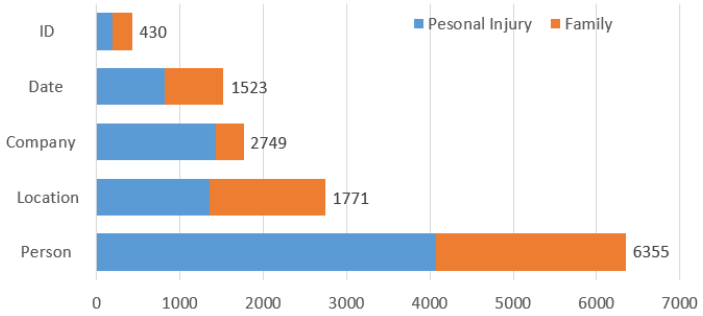


Figure 1. Entity type distribution on 118 legal documents

3.1. Annotation phase

The Doccano tool (Nakayama et al, 2018) was configured to allow quick and easy manual annotation of text with the 5 predefined entities. Experts found around 13.000 entities as shown in the figure 1. In a personal injury case, there are many more individuals involved (the parties, their insurance companies, businesses, doctors, witnesses, and their attorneys) than in a divorce or separation case, which generally involves only family members.

Finally, the dataset was split 80% for training models and 20% for model evaluation in a stratified fashion to preserve the distribution of entities across both subsets, as shown in Table 2.

Table 2. Entity type distribution on train and test set

Entity	Train set	Test set
Person	5287	1068
Location	2274	475
Company	1490	281
Date	1254	269
ID	355	75

4. Model Architectures

The entity recognizer models were trained using spaCy v3.2's Named Entity Recognition system (Honnibal and Montani, 2017). It uses transition-based Chunking Model introduced by Lample et al. 2016, where the task of predicting the structure of the text is mapped to a series of state transitions. During training, it is parameterized by defining a probability distribution over the actions at each time step, given the current content of the stack, buffer, and output, as well as the history of actions taken. This distribution is used to predict the next action given the current state.

The complete pipeline consists of three subnetworks:

1. A model to map each token into a context-dependent representation.
2. A model to encode the states by summing the component features (words processed and to be processed) and applying the non-linearity.
3. A feed-forward network that predicts actions from the state representation. The final prediction of the actions on the token of the text can be done through beam-search (Zhang and Clark, 2011).

We have focused on point 1 because word representations are critically important in training effective entity recognition models. The standard spaCy's contextual encoder consists of a 3-gram convolution layer. It maps the vector representations of a token and its two neighbors into a lower-dimensional output vector. It uses word vectors trained using FastText CBOW on Italian Text (from Wikipedia and OSCAR, oscar-corpus.com).

Instead, we use two Transformer-based models pre-trained on large Italian corpora, Italian XXL BERT and Italian XXL ELECTRA, available on Huggingface model hub (Wolf et al, 2020) and trained by the MDZ Digital Library team at the Bavarian State Library. They are able to better capture the context of each word in text than other classic word embedding models since they can preserve all the information of the entire sequence fed into the model. They can differentially weight the significance of each part of the input data according to their context. The Italian XXL BERT model has the Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) architecture and Italian XXL ELECTRA has the Efficiently Learning and Encoder that Classifies Token Replacement Accurately (ELECTRA, Clark et al., 2020) architecture. Both models have been trained on articles from Wikipedia in Italian, various texts from the OPUS corpora collection (opus.nlpl.eu), and data from the Italian part of the OSCAR corpus (oscar-corpus.com).

5. NER Models Results

Model performance was evaluated on the RECALL metric to place more emphasis on errors on false negatives and identify as many personal identifiers as possible. Then, NER models were optimized to maximize RECALL on a

validation set consisting of 10% of the training data split in a stratified fashion. These data were excluded from the training phase of the model but they were used to monitor the learning process and maximize RECALL metric. Table 3 reports our comparisons between the transition-based NER model using the large Italian spaCy's contextual encoder (Tok2vec_ITA_NER), the Italian XXL ELECTRA (ELECTRA_ITA_NER), and the Italian XXL BERT (BERT_ITA_NER) pretrained embeddings.

Table 3. It shows performance comparisons between Tok2vec_ITA_NER, BERT_ITA_NER and ELECTRA_ITA_NER models on individual entities. TP and FN refer to True Positives and False Negatives, respectively. Recall is calculated from the formula TP/(TP+FN)

Type	Tok2vec_ITA_NER			BERT_ITA_NER			ELECTRA_ITA_NER		
	TP	FN	Recall %	TP	FN	Recall %	TP	FN	Recall %
Company	156	124	55.7	230	50	**82.1**	220	60	78.6
Date	173	96	64.3	239	30	88.8	251	18	**93.3**
ID	53	22	70.7	73	2	97.3	75	0	**100**
Location	437	38	92	452	23	**95.2**	446	29	93.9
Person	864	203	81	1058	9	**99.2**	1059	8	99.3
OVERALL	1683	483	77.7	2052	114	**94.7**	2051	115	**94.7**

As we can see in table 3, BERT_ITA_NER and Electra_ITA_NER have similar performance with 94.7% Recall, but Date and ID entities are better identified by the ELECTRA-based model.

6. Anonymization Type

The anonymization system we have implemented supports different anonymization methods that can be applied to the text depending on the purpose of the anonymization:

Suppression: replacing named entities (sensitive information) with a fixed value (e.g. "Daniele lives in Milan" -> "OMISSIS lives in OMISSIS")

Initials: replacing named entities with the initials (e.g. "Daniele lives in Milan" -> "D. lives in M.")

Tagging: replacing named entities with the entity name (e.g. "Daniele lives in Milan" -> "PERSON lives in LOCATION)

Tagging Context Preserving: replacing named entities with the entity name and occurrence of the corresponding entity type (e.g. "Daniele lives in Milan. Daniele and Mario are .." -> "PERSON1 lives in LOCATION1. PERSON1 and PERSON2 are ..")

Tagging Generalization: similar to Tagging Context Preserving but it identifies

a text entity that refers to an object of the same class (e.g. Daniele Licari, Daniele o LICARI -> PERSON1). It checks if a text entity is a subset of others found in the text (e.g.“Daniele Licari lives in Milan. Licari and Mario are ..” -> “PERSON1 lives in LOCATION1. PERSON1 and PERSON2 are..”)
Random Substitution Context Preserving: replacing named entities with a random entity of the same class using a default list containing random entities of each class (e.g. “Daniele lives in Milan. Daniele and Mario are …” -> “Carlo lives in Rome, Carlo and Giovanni are ..”)

7. Anonymization Service

The text anonymization service is offered through a web application (shown in Fig 2). Users can enter the text to be anonymized or upload a PDF document, select the NER model, the entities to be found, and the anonymization method to be used. Once the input data is provided, the system preprocesses it (e.g. converting the pdf to text and minimizing conversion errors), applies the selected NER model (BERT_ITA_NER or ELECTRA_ITA_NER) and the selected anonymization method to obtain a shareable anonymized version. The application was developed using Streamlit framework.

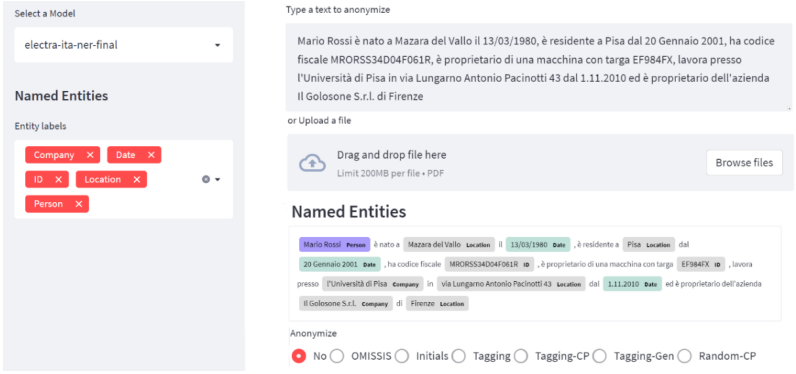


Figure 2. Screenshot of the Anonymization Service

8. Conclusion

In this work, we have designed a service prototype for a fully automated anonymization system of the Italian legal text. The system can automatically convert the PDF document into raw text, recognize the entities to be anonymized and propose different anonymization strategies (e.g. Suppression, Tagging, Random Substitution).
Our models are based on pre-trained Transformers embeddings and SpaCy transition-based parsing for entity recognition. We obtained high recall scores

on single entities, despite the pessimistic estimate due to labeling errors (in the selection of entity boundaries), especially on companies. The model was trained to anonymize judgments on personal injury damage and family from the Court of Genoa. These subjects were chosen based on our team's long experience in identifying personal data on this type of data and because of the scientific collaboration agreement with the Court of Genoa.

Our experiments suggest that it is possible to automate the anonymization of Italian legal text using deep neural network models. Our first application domain was the court of Genoa and the service is in the testing phase. This is the first step in creating a privacy-friendly (GDPR-compliant) Italian legal data publishing system.

In the future, we will test deep learning models in combination with rule-based expert systems (regular expressions and/or dictionaries) to increase overall performance. Post-processing can also improve the Recall of predictions. For example, combining the results of the two models BERT_ITA_NER and Electra_ITA_NER or ensuring that the named entities found do not appear elsewhere in the text without being tagged (Guillarme and Thuiller, 2021). Additionally, we will expand the analysis to other legal subjects and other legal databases after signing a collaboration agreement with the courts. This research aims to become the information source to support the development of innovative platforms for the investigation and interpretation of legal documents, such as predictivejustice.eu.

References

Barriere V. and Fouret A. (2019). May I Check Again? — A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*.

Clark K., Luong M. T., Le Q. V. and Manning C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Devlin J., Chang M. W., Lee K. and Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garat D. and Wonsever D. (2022). Automatic Curation of Court Documents: Anonymizing Personal Data. *Information*, *13*(1), 27.

Glaser I., Schamberger T. and Matthes F. (2021, June). Anonymization of german legal court rulings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 205-209).

Honnibal M. and Montani I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, *7*(1), 411-420.

Kleinberg B., Mozes M., van der Toolen Y. and Verschuere B. (2017, June 3). NETANOS - Named entity-based Text Anonymization for Open Science. https://doi.org/10.31219/osf.io/w9nhb.

Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.
Le Guillarme N. and Thuiller W. (2022). TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, *13*(3), 625-641.
Mamede N., Baptista J. and Dias F. (2016, July). Automated anonymization of text documents. In *2016 IEEE congress on evolutionary computation (CEC)* (pp. 1287-1294). IEEE.
Nakayama H., Kubo T., Kamura J., Taniguchi Y. and Liang X. (2018). doccano: Text annotation tool for human. *Software available from https://github. com/doccano/doccano*.
Oksanen A., Tamper M., Tuominen J., Hietanen A. and Hyvönen E. (2019, September). ANOPPI: A Pseudonymization Service for Finnish Court Documents. In *JURIX* (pp. 251-254).
Oleksy M., Ropiak N. and Walkowiak T. (2021). Automated anonymization of text documents in Polish. *Procedia Computer Science*, *192*, 1323-1333.
Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi, A., Cistac P., Rault T., Louf R., Funtowicz M. and Brew J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
Zhang Y. and Clark S. (2011). Syntactic processing using the generalized perceptron and beam search. *Computational linguistics*, *37*(1), 105-151.

Text Mining and Variants of Correspondence Analysis for analysing written Italian of University students

Rosaria Lombardo[1], Matteo Pascoli[2], Nicola Grandi[2]

[1]University of Campania "L. Vanvitelli" –
rosaria.lombardo@unicampania.it
[2]University of Bologna – matteo.pascoli2@unibo.it
nicola.grandi@unibo.it

Abstract
This paper aims at investigating the textual similarities/dissimilarities of the written Italian used by university students of different departments of North, Centre and South Italy. The students' text data are part of a large survey dataset, collected from the University of Bologna aiming at analysing an alleged decline in Italian language and at highlighting peculiar linguistic features of the Italian language used by university students.
The text data comes from a sample of 2159 participants belonging to different departments of Italian universities.
Here we focus on studying the association between the written production of semiformal texts of Italian students and the University geographical area (North, Centre and South of Italy) through a non-symmetrical variant of simple correspondence analysis.

Keywords: Text-Mining, Linguistic Variation, Variants of Correspondence Analysis, Statistically Significant Association.

Riassunto
Questo contributo si propone di indagare le somiglianze/dissomiglianze testuali dell'italiano scritto utilizzato da studenti universitari di diversi dipartimenti del Nord, Centro e Sud Italia. I dati testuali degli studenti fanno parte di un ampio dataset di indagine, raccolto dall'Università di Bologna con l'obiettivo di analizzare un presunto declino della lingua italiana e di mettere in evidenza le caratteristiche linguistiche peculiari della lingua italiana utilizzata dagli studenti universitari.
I dati testuali provengono da un campione di 2159 partecipanti appartenenti a diversi dipartimenti di atenei italiani.
Qui ci concentriamo sullo studio dell'associazione tra la produzione scritta di testi semiformali di studenti italiani e l'area geografica dell'Ateneo degli studenti (Nord, Centro e Sud Italia) attraverso una variante non-simmetrica dell'analisi semplice delle corrispondenze.

Parole Chiave: Text-Mining, Variazioni Linguistiche, Varianti dell'Analisi delle Corrispondenze, Associazione Statisticamente Significativa.

1. Introduction

In literature (Telmon, 1993; Cerruti, 2011) some language differences between the North, Centre and the South of Italy are well known. When Italian people speak, not only the accent, but also some words that they use result different. For example, it is well established that some Italian words are more used in North than in South Italy. Here, we wonder whether among young people the prevalent use of some *words* is still common, so words like «Essere vs Stare, Avere vs Tenere, Nulla vs Niente, Anche vs Pure» will be taken into account and counted in the Italian texts of the university students who took part to the survey promoted by the University of Bologna (Italy). Here, we investigate if such language differences still result significant in the Italian students' written texts. In particular, we considered the same number of written texts for students of North, Centre and South of Italy. In 2020, the Italian students were asked to write a text piece of no more 500 words about their experience of «didactic on line» during the pandemic period. By using text mining tools, we counted the presence of a selected list of words which could highlight some language difference across Italy.

To analyse and visualise the significant association between *words* and *geographical area* (North, Centre and South), we consider a variant of correspondence analysis, i.e. non-symmetrical correspondence analysis (Lauro and D'Ambra, 1984; D'Ambra and Lauro, 1989; Kroonenberg and Lombardo, 1999) where the *geographical area* has been seen as the predictor variable of the *words* used by the university students. Also, for each variable category has been computed an approximate p-value (Beh and Lombardo, 2015) to assess the statistical significance of the non-symmetric association.

2. A Variant of Correspondence Analysis

Non-symmetrical correspondence analysis (NSCA) is a variant of simple correspondence analysis and is implemented in the `R` package CAvariants (Lombardo and Beh, 2016) together with other five variants. NSCA is a method increasingly used in place of simple correspondence analysis when the underlying structure of the association between the row and column variables is assumed to be asymmetric (Lauro and D'Ambra, 1984; D'Ambra and Lauro, 1989). Doing so, a column variable has been treated as a predictor categorical variable and a row variable as a response categorical variable. Here, the *geographical area*, (with three categories, North, Centre and South) has been assumed useful to explain the students' use of some Italian *words* (with 11 categories given by some common words described in the next section) across Italy. NSCA is based on the singular value decomposition of the predictability

index tau (Goodman and Kruskal, 1954). As a result, a visualisation of the asymmetric association between the categorical variables and a statistical test for detecting significant association between categories will be of interest. Consider an IJ two-way contingency table, $\mathbf{N}$, where the (i, j)'th cell entry is given by for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$.

Denote the grand total of $\mathbf{N}$ by n and the (i, j)'th relative frequency by $p_{ij} = n_{ij} / n$.
Define the i'th row relative marginal frequency by

$$p_{i\bullet} = \sum_{j=1}^{J} p_{ij} \tag{1}$$

and the j'th column relative marginal frequency by

$$p_{\bullet j} = \sum_{i=1}^{I} p_{ij} \tag{2}$$

The Goodman-Kruskal tau index, τ, on which NSCA rests, is given by

$$\tau = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2 \Bigg/ 1 - \sum_{i=1}^{I} p_{i\bullet} \tag{3}$$

A more formal statistic than can be used to test the significance of the predictability can be made by considering the C-statistic of Light and Margolin (1962) which follows a chi-square distribution with (I-1)(J-1) degree of freedom

$$C = (n-1)(I-1)\tau \sim \chi^2_{(I-1)(J-1)} \tag{4}$$

For further details see, for example, Beh and Lombardo (2014, chap.5). Broadly speaking, the C-statistic allows to test the dependence between the two variables, but to know when a category becomes a statistically significant contributor to the predictability structure one should use the approximate p-value of a point category related to the confidence ellipse of each categorical point (Beh and Lombardo, 2015). For the j'th column predictor category, the p-value is given by

$$p - value = P\left\{\chi^2 > (n-1)(I-1)\tau\left[\left(\frac{g_{j1}}{\lambda_1}\right)^2 + \left(\frac{g_{j2}}{\lambda_2}\right)^2\right]\right\} \tag{5}$$

where λ_1 and λ_2 are the first two singular values of the τ decomposition, g_{j1} and g_{j2} are the principal component coordinates of the j'th column (predictor) category along the first and second axis of a correspondence plot. Similarly, expressions for the calculation of p-values for the row (response) categories may be derived. When visualizing results, these p-values allow to assess the significance of the proximity of a categorical point from other points in the same plot and the proximity of a point from the origin of the plot, which coincides with where all the points would be if there were complete independence between the row and column variables.

3. Application

In this application, we consider the text data of Table 1 which are part of a large survey dataset, collected from the University of Bologna in 2021 (Prot 2017LAP429), aiming at analysing an alleged decline in Italian language and at highlighting peculiar linguistic features of the Italian language used by university students. The text data comes from a sample of 2159 participants belonging to different departments of Italian universities, who were asked to fill in a questionnaire and to write a text whose subject was related to the online didactic experienced during the pandemic year. By using the R package, tm, for text mining (Feinerer, Hornik and Meyer, 2008), considering the same number of written texts (573) of university students for each geographical area, one can look at the *words* more frequent in those semiformal texts (see Figure 1, larger

Figure 1. Students' words cloud of the most frequent words

words correspond to more counted words). Also, considering a list of suitable *words* which highlight some possible language difference across Italy (Telmon, 1993; Cerruti, 2011; Thomason and Kaufman, 1988), we made the contingency table **N** given in Table 1, where the row variable categories are given by 11 *words* and the column variable categories are given by the geographical areas. The eleven Italian *words* considered are *Essere*, *Stare*, *Avere*, *Tenere*, *Nulla*, *Niente*, *Anche*, *Casa*, *Tempo*, *Lavoro and Dad* (that is an Italian acronym for "distance learning") while the geographical areas are North (N), Centre (C) and South (S) of Italy.

Table 1. Cross-classification of the key-words and the geographical area of students.

	N	***C***	***S***	*Tot*
Essere	647	690	617	1954
Stare	127	145	143	415
Avere	282	300	310	892
Tenere	42	35	27	104
Nulla	31	29	39	99
Niente	23	33	32	88
Anche	54	66	64	184
Casa	503	479	455	1437
Tempo	685	593	448	1726
Lavor	323	288	188	799
Dad	344	469	524	1337
Tot	3061	3127	2847	9035

Table 2. approximate p-values of students'words

	HL Axis1	HL Axis2	P-value
Essere	0.017	0.003	0.001
Stare	0.017	0.003	1.000
Avere	0.017	0.003	0.165
Tenere	0.017	0.003	1.000
Nulla	0.017	0.003	0.995
Niente	0.017	0.003	1.000
Anche	0.017	0.003	1.000
Casa	0.017	0.003	0.000
Tempo	0.017	0.003	0.000
Lavor	0.017	0.003	0.000
Dad	0.017	0.003	0.000

Table 3. approximate p-values of geographical areas

	HL Axis1	HL Axis2	P-value
N	0.029	0.004	0.000
C	0.029	0.004	0.000
S	0.030	0.005	0.000

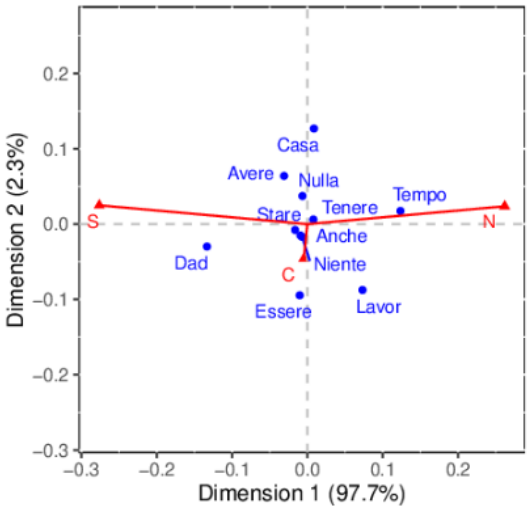


Figure 2. Column Biplot of Table 1

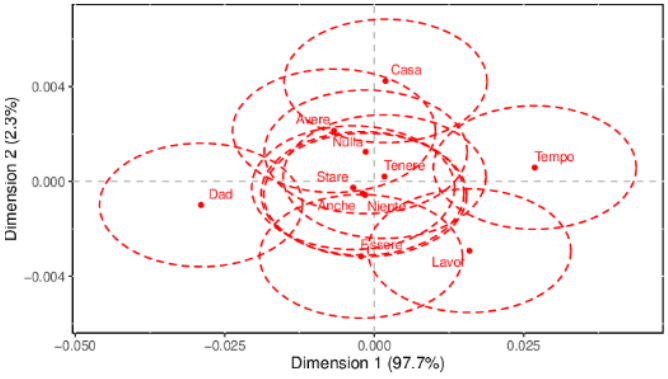


Figure 3. Confidence Ellipses at 95% superimposed on the 11 words considered

One may consider that the association between *words* and *geographical areas* be more naturally asymmetrically structured such that the column variable (*geographical area*) is treated as a predictor variable and the row variable (common *words*) is treated as a response variable.

In doing so, we perform a non-symmetric correspondence analysis. The Goodman-Kruskal tau index is but its related C–statistic is 205.98 (p-value < 0.0001), therefore, there is a highly statistically significant asymmetric association between the geographical area (predictor) and the words (response) considered. Thus, the geographical areas of students' University do impact upon how the Italian words used in the Italian written texts. A two-dimensional visual summary of the asymmetric association can be found by considering Figure 2. The first dimension accounts for 97.7% of the asymmetric association that exists

between the two variables, and the second dimension for only 2.3%. Therefore, the first two dimensions account for 100% of the asymmetric association. Also, Figure 2 is an optimal biplot where the predictors are depicted as vectors and the responses as points. The closeness of a point to the vector gives information on the strength of the prediction, however for the interpretation we focus our attention only on those *words* (responses) and *geographical areas* (predictors) which show a p-value less than 0.001 (see Tables 2 and 3). Tables 2 and 3 provide the length of the semi-axis of these confidence ellipses and the related p-values for the considered *words* and the *geographical areas*. Table 2 shows that the p-values of some *words*, i.e. *Stare*, *Avere*, *Tenere*, *Nulla*, *Niente*, and *Anche*, are higher than 0.05, therefore these *words* are not significantly predicted. Differently, the p-values of *Essere*, *Casa*, *Lavoro*, *Tempo* and *Dad* are less than 0.001, therefore these words have been impacted by the geographical area of students' University. Table 3 shows that the p-values associated with the three geographical areas are less than 0.001, so all predictor categories are statistically significant. Portraying the asymmetric association, Figure 2 shows a clear difference between the South and the North of Italy, given the opposition of the vector depicting the South (S) with that of North (N). In South Italy the *word* more used is *Dad*, while in the North Italy the more frequent *words* are *Tempo* and *Lavoro*, but *Lavoro* is also associated with the Centre Italy. The word *Casa* that is not used so much in the Centre (because its position is opposite to the vector depicting the Centre) results used in the South and North Italy in the similar way. Note that the word *Essere* is strongly associated with the Centre with respect to the North and South Italy where it seems similarly used, disappointing previous study on the language geographical differences (Telmon, 1993). Finally, Figure 3 shows the 95% confidence ellipses for each response categories. Those *words* whose ellipse does not include the origin of the plot are significantly predicted.

4. Conclusion

This paper has examined procedures for quantifying and interpreting the written production of semiformal texts of Italian students, taking into account the geographical area of students' University (North, Centre and South of Italy), through text mining techniques and a non-symmetrical variant of classical correspondence analysis. As a result, among young people of the North and South Italy, the use of some *words* like «Essere vs Stare» seems to be changed, and some language differences do not result significant in the Italian students' written texts. Our application has been focused on a simple two-way contingency table, but more or different variables might be considered to describe peculiar linguistic features of the Italian language and other variants of correspondence analysis could be used. However, we shall leave these issues for future investigation.

References

Beh E. J. and Lombardo R. (2014a). *Correspondence Analysis, Theory, Practice and New Strategies*. John Wiley & Sons, Chichester, UK.

Beh E. J. and Lombardo R. (2015b). Confidence regions and Approximate P-values for classical and non-symmetric correspondence analysis. *Journal of Communications and Statistics*, Theory and Methods, vol. 44: 95–114.

Cerruti M. (2011). Regional varieties of Italian in the linguistic repertoire. *International Journal of the Sociology of Language*, vol. 210: 9—28.

D'Ambra L. and Lauro N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency table. In: Coppi, R., Bolasco, S., eds. *Multiway Data Analysis* (pp. 301–315). Amsterdam: Elsevier.

Feinerer I., Hornik K. and Meyer D. (2008). Text mining infrastructure in R. *Journal of statistical software*, 25, 1-54.

Greenacre M. J. (1984). *Theory and Application of Correspondence Analysis*. Academic Press, London.

Lauro N. C. and D'Ambra L. (1984). L'Analyse non symmétrique des correspondances. In E. Diday (Ed.), *Data analysis and informatics III* (pp. 433–446). Amsterdam: Elsevier.

Light R. J. and Margolin B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, vol. 66: 534–544.

Lombardo R. Beh E. J. (2016). Variants of Correspondence Analysis. *The R Journal*, vol. 8 (2): 167—184.

Kroonenberg P. M. and Lombardo R. (1999). Non-symmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, vol. 34: 367–396.

Telmon T. (1993). Varietà regionali, in Alberto A. Sobrero (a cura di). *Introduzione all'italiano contemporaneo. La variazione e gli usi* (pp. 93-149). Roma-Bari, Laterza,

Thomason S. and Kaufman T. (1988). *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Corpus non alignés et ADT. Essai de comparaison entre les présidents français et brésiliens de l'ère contemporaine

Carlos Maciel, Damon Mayaffre, Laurent Vanni

CNRS, UCA, BCL – Carlos.Maciel@univ-cotedazur.fr
Damon.Mayaffre@univ-cotedazur.fr Laurent.Vanni@univ-cotedazur.fr

Abstract
Is there an ADT method that can deal with non-aligned bilingual corpora? Does the textual genre exert a sufficiently strong constraint on the discourse that would make texts written in different languages comparable, provided they are of identical genre? To answer these two questions, one methodological, the other linguistic, this contribution gathers in a single corpus French and Brazilian presidential speeches of the contemporary era (1950-2020), from de Gaulle to Macron, from Kubitschek to Lula, i.e. 15 million words. A methodological path is proposed from the simple frequency dictionary to the factorial treatment of the cooccurrencial profiles of words, in order to establish a generic transnational presidential speech.

Keywords: presidential speech, logometry, textual genre, isotopy, correlates, co-occurrences.

Résumé
Existe-t-il une méthode ADT susceptible de traiter des corpus bilingues non alignés ? Le genre textuel exerce-t-il une contrainte suffisamment forte sur le discours qui rendrait comparable des textes écrits dans des langues différentes sous condition d'être de genre identique ? Pour répondre à ces deux questions, une méthodologique, l'autre linguistique, cette contribution rassemble dans un même corpus les discours politiques présidentiels français et brésiliens de l'époque contemporaine (1950-2020), de de Gaulle à Macron, de Kubitschek à Lula, soit 15 millions de mots. Un parcours méthodologique est proposé du simple dictionnaire des fréquences jusqu'au traitement factoriel des profils concurrentiels des mots, afin d'établir un parler présidentiel générique transnational.

Mots clés : discours présidentiel, logométrie, genre textuel, isotopie, corrélats, cooccurrences.

1. Introduction

L'ADT (Iezzi et al., 2020, Lebart et Salem, 1994 ; Lebart, Pincemin et Poudat, 2019) analyse depuis plusieurs décennies principalement des *corpus contrastifs*

monolingues. Le corpus de textes, partitionné en sous-corpus selon une certaine métadonnée (l'auteur, la date, le genre…), constitue alors la référence ou la norme statistico-linguistique, par rapport à laquelle les sous-corpus en question peuvent se singulariser (calcul de spécificités, calcul de distances intertextuelles, etc.). Différemment, (Zimina, 2000), aux JADTs de Lausanne, puis dans diverses publications, a proposé une méthodologie ADT, implémentée dans le logiciel *Lexico*, pour étudier statistiquement des *corpus bilingues alignés*, répandus dans le TAL (Véronique *éd*., 2000) : il s'agissait alors de comparer grâce aux outils textométriques des textes jumeaux et parallélisés – de strictes traductions – écrits dans des langues différentes, par exemple les textes officiels canadiens rédigés en français et en anglais.

Soit, ici, aujourd'hui, ni un corpus contrastif monolingue, ni un corpus bilingue aligné, mais un corpus composé de textes que l'analyste juge comparables, et pourtant écrits dans des langues différentes, par des locuteurs différents et dans des situations distinctes : *un corpus bilingue non parallélisable*. Que peut en dire l'ADT ?

Les discours des présidents de la République français et des présidents de la République brésiliens ont été ainsi rassemblés sur une longue période historique 1950-2020 dans un corpus français/portugais non aligné de 15 millions de mots[1]. L'hypothèse linguistique forte que nous mettons à l'épreuve est empruntée à (Brunet, 2009), dans la lignée historique des travaux de Volochinov ou Bakhtine : le genre discursif – ici le genre « discours présidentiel » – n'est pas seulement un pallier intermédiaire entre la langue et le discours, mais un pallier transversal aux langues, qui relie les discours. Autrement dit, le genre contraint ou informe les productions langagières, par-delà les langues, et rend comparable la production discursive d'un Macron ou d'un Kubitschek avec celle d'un de Gaulle ou d'un Lula[2].

A titre introductif, nous concédons que (1) la comparabilité problématique que nous éprouvons, au titre d'un genre commun, grâce à l'approche quantitative, n'apparait envisageable que par la proximité étymologique ou culturelle des deux langues envisagées (le français et le portugais brésilien) ; que (2) les méthodes ADT que nous proposons peinent à être strictement formalisées : il s'agit d'avantage d'un parcours méthodologique « bricolé » au sens de Levy Strauss, convoquant l'AFC, le traitement de la cooccurrence ou le simple dictionnaire de

[1] Le sous-corpus français rassemble les grands textes de de Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy, Hollande et Macron, entre 1958 et 2021, équivalant 4,5 millions de mots. Le sous-corpus brésilien, plus important en taille, rassemble la plupart des discours de Kubitschek, Goulart, Castelo Branco, Médici, Geisel, Figueiredo, Sarney, Collor, Franco, Cardoso, Lula, entre 1956 et 2010, équivalant 11 millions de mots.

[2] Précisons encore que le Brésil et la France ont des régimes présidentiel et semi-présidentiel semblables qui donnent à la parole présidentielle française et brésilienne une autorité identique.

fréquences que d'une méthode intégrée.

2. La contrainte du genre : un stock lexical commun

Si les calculs aussi éprouvés sur corpus contrastif monolingue que le *z-score* ou les *spécificités* apparaissent impuissants pour traiter nos données bilingues non alignées, un simple détour par les dictionnaires de fréquence respectifs des sous-corpus français et brésilien apparait instructif des contraintes génériques qui pèsent sur les discours.

Après étiquetage et lemmatisation[3], le logiciel *Hyperbase* sélectionne hiérarchiquement les 20 substantifs les plus utilisés, respectivement par les présidents français et les présidents brésiliens. Or 18 sur 20 apparaissent communs aux deux ensembles, parfois au rang près d'utilisation (figure 1).

Ainsi, par exemple, nos présidents placent au premier rang 'pays'/'país', et au deuxième rang 'année'/'ano'. Plus loin dans le tableau, 'chose', 'problèmes', 'monde', etc. en français, correspondent strictement en portugais à 'coisa', 'problema', 'mundo' etc.

Il existe donc un stock lexical fortement contraint du discours présidentiel que l'on pourrait qualifier d'épilexique générique, indépendant de la langue et du pays. Et de part et d'autre de l'Atlantique, lorsqu'on est président de la République, on parle à ses concitoyens de la 'vie'/'vida', du 'travail'/'trabalho', du 'temps/'tempo', de 'président'/'presidente', etc.

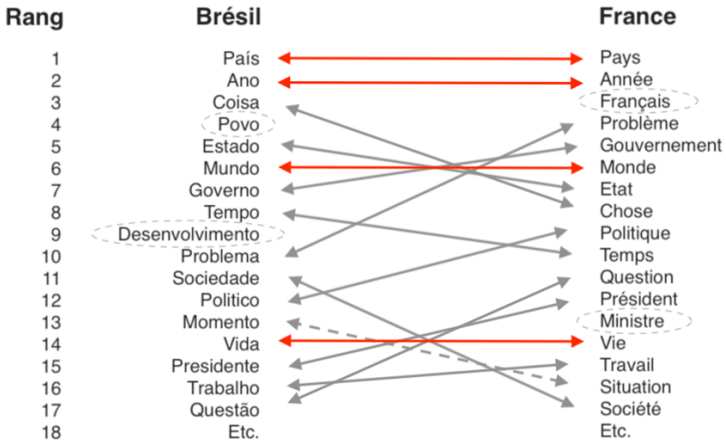


Figure 1. Etude comparée du dictionnaire de fréquences (Brésil/France). Substantifs classés par leur rang de fréquences

[3] Le corpus a été lemmatisé et étiqueté morphosyntaxiquement par les logiciels Spacy et TreeTagger dont on sait les limites de performance mais qui ont l'avantage de fonctionner sur le français et le portugais.

Dans cette stéréotypie lexicale des discours présidentiels que révèle le tableau des fréquences, il deviendrait important de commenter les très rares exceptions, puisqu'elles soulignent des particularismes nationaux France/Brésil suffisamment forts pour résister au genre présidentiel. Par exemple, les acteurs principaux des discours en France sont les 'Français' (3ème rang), là où le discours brésilien ne convoque pas les 'Brésiliens' et semble s'adresser avant tout au peuple ('povo', 4ème rang). Il y aurait ainsi un patriotisme mieux ancré dans les habitudes lexicales au sein du vieux cadre national français, là où la jeune République fédérale se sert d'un vocabulaire plus politique que patriotique. Par ailleurs, le terme de 'desenvolvimento' très haut placé au Brésil (9ème rang) n'a pas d'équivalent dans les 20 premiers substantifs en France ('développement') et marque sans doute, sur la longue durée, la problématique d'un pays en pleine expansion politique, démographique, économique, sociale durant la période, là où la France post-30 glorieuses et post-coloniale se vit peut-être davantage sur le déclin ou la préservation plutôt que dans le développement.

3. La contrainte du politique : des thématiques croisées

Des mots aux thèmes, l'ADT propose depuis (Benzécri, 1973) dans son principe fondamental puis (Viprey, 1997) dans une application circonstanciée, l'étude des thématiques du discours – appelées *isotropies* et implémentées dans le logiciel *Hyperbase* sous le nom de *Corrélats* : seraient-elles identiques ou voisines dans les sous-corpus français et brésiliens ?

Pour chaque pays, nous construisons une matrice de contingence carré (ou triangulaire puisque symétrique à la diagonale) composée en ligne des 300 substantifs les plus utilisés du discours (dont les 20 mentionnés plus haut) et en colonne ces mêmes substantifs ; dans les cellules du tableau, le nombre de fois où les mots considérés cooccurrent. Le traitement factoriel des profils coocurrentiels des mots ainsi constitués donne les deux représentations suivantes. S'il nous apparait hasardeux mathématiquement de prétendre à une comparaison formelle des deux sorties machines[4], l'exploration visuelle des deux nuages permet de conclure à des ressemblances évidentes, sans doute imputables au genre commun.

Le nuage de mots français se lit facilement en 4 isotropies majeures selon des résultats que nous avons souvent décrits et analysés (Mayaffre, 2012). Le discours présidentiel français a ainsi une isotropie ou composante (1) institutionnelle ('majorité', 'assemblée', 'mandat', etc.), une composante (2)

[4] Par exemple, le fait que la contribution aux axes 1 et 2 se chiffrent respectivement à 6,1% et 3,2% dans de corpus français et à 8,5 et 6,3% dans le corpus brésilien nous parait difficilement interprétable sans prendre de gros risques sur-interprétatifs.

économique (‘déficit’, ‘prix’, ‘banque’, ‘agriculture’, etc.), une composante (3) sociale et sociétale (‘chômage’, ‘retraite’, ‘contrat’ mais aussi ‘enfant’, ‘éducation’, ‘famille’) et une composante (4) internationale (‘arme’, ‘paix’, ‘guerre’, ‘conflit’, ‘puissance’, etc.).

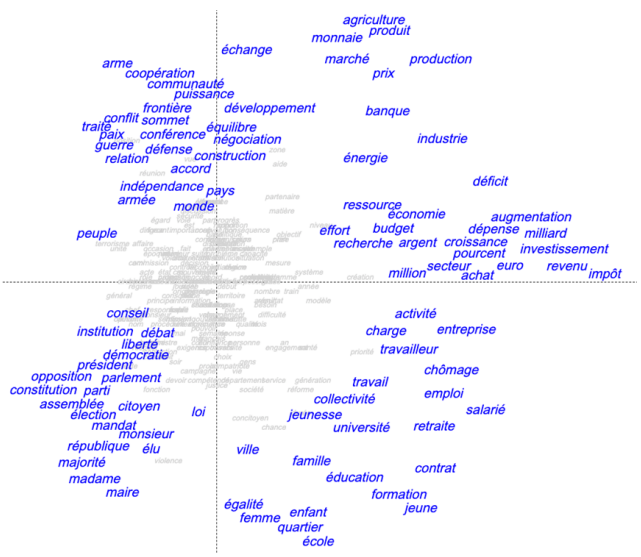


Figure 2. Corrélats des 300 substantifs le plus utilisés dans le corpus français (Hyperbase – 2022)

Dans le discours brésilien, nous semblons retrouver l’essentiel de cette information thématique – même si nous soulignerons les nuances. Nous repérons ainsi facilement l’isotropie (1) institutionnelle (‘ministro’, ‘governador’, ‘deputado’ ‘prefeito’), l’isotropie (2) économique (‘dólar’, ‘taxa’, ‘inflação’), et l’isotropie (3) sociale et sociétale avec ‘emprego’, ‘salário’, ‘pobre’, ‘família’, ‘escola’ ou ‘universidade’ etc. Les mots peuvent différer entre les deux pays, mais la structuration thématique ou isotropique des textes du corpus apparait identique.

Cependant l’isotropie de politique internationale très prégnante en France ne semble pas organiser le discours présidentiel brésilien, comme il structure le discours Français. La quatrième composante brésilienne compte ainsi des termes moins géopolitiques et plus politiques qu’en France (comme nous l’avions pressenti avec ‘povo’ vs. ‘Français’). Ainsi, par exemple, au ‘conflit’ en France semble répondre l’‘integração’ au Brésil, à la ‘puissance’ française semble répondre le ‘desenvolvimento’ brésilien. Là où en France nous enregistrons ‘guerre’ ou ‘arme’, etc., nous enregistrons au Brésil ‘liberdade’, ‘democracia’,

etc. Cette différence ne semble pas devoir surprendre au regard de l'histoire et de la position comparées du Brésil et de la France dans le concert international. La France fait partie des puissances nucléaires, le Brésil non. La France fait partie de manière permanente du Conseil de sécurité, le Brésil non. La France est un ancien pays colonial qui garde une forte influence en Afrique, le Brésil non. Du reste, d'autres facteurs explicatifs pourraient être avancés parmi lesquels les prérogatives du président accordées par les Constitutions respectives (chef des armées en France comme l'écrit hautement l'article 15 de la Constitution, moins revendiqué au Brésil) et par la culture politique différente (le Brésil ne fait pas de la guerre un point central de son action internationale, là où l'affirmation de la puissance française sur les différents continents est une constante des politiques françaises aux XX^{ème} et XXI^{ème} siècle.).

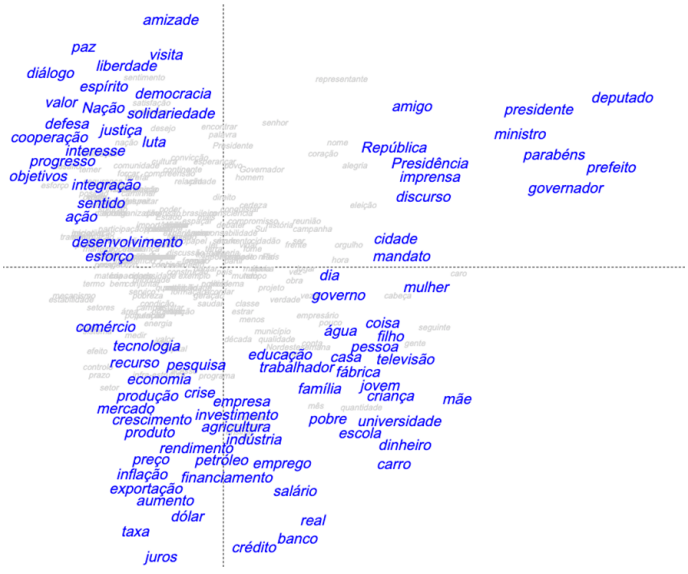


Figure 3. Corrélats des 300 substantifs les plus utilisés dans le corpus brésilien (Hyperbase 2022)

4. Des thèmes au sens : les limites d'une comparaison multilingue

Des mots aux thèmes, des thèmes au sens, nous proposons pour conclure de zoomer sur l'utilisation respective d'un terme partagé et très utilisé dans le deux sous-corpus : 'travail' / 'trabalho'. Nous rappelons avec la linguistique contextuelle anglo-saxonne (par ex. Halliday et Hasan, 1976) et française (par ex. Guiraud, 1960) que le sens d'un mot est la somme de ses contextes d'utilisation. Et précisons avec (Mayaffre, 2014) que cette affirmation signifie,

statistiquement, que le sens d'un mot peut être considéré comme *la somme de ses cooccurrences*.

Partant, le nuage des cooccurrents du mot 'travail'/'trabalho' laisse paraitre des nuances sémantiques fortes entre le France et le Brésil (figure 4). Certes on remarquera quelques mots cooccurrents identiques 'fruit'/'fruto', 'travailleur'/'trabalhador', 'emploi'/'emprego' ou d'une certaine manière 'entreprise'/'oficina'.

Néanmoins, le sens de travail apparait très incarné au Brésil. Il s'adresse aux familles ('família'), aux enfants ('criança'), aux hommes et aux femmes ('homem' et 'mulher') ou encore à la main d'œuvre ('mao-de-obra') ou aux équipes ('equipe'). Le travail au Brésil se définit ainsi avant tout comme le produit d'une activité humaine. En France les acteurs du travail semblent avoir disparu du discours (à l'exception de 'travailleur'), et la question du travail se technicise autour de revendications sociales historiques ('cotisation', 'branche', 'salaire', 'réforme', etc.). En France, le travail est avant tout regardé comme une organisation économique et sociale. De manière plus obvie, le travail relève au Brésil, pays où le secteur agro-industriel tient une place dominante dans l'économie, du travail des champs ('campo', 'cana', 'canavial', 'sisal') et soulève la question essentielle du partage des terres ('terra').

Figure 4. Cooccurrents de 'travail'/'trabolho' (la taille des mots est proportionnelle à l'indice de cooccurrence)

4. Conclusion

La comparaison de corpus bilingues non alignés ne semble pas pouvoir être automatisée. Le passage par une langue-pivot est inenvisageable puisqu'il abolirait les nuances sémantiques que le linguiste se propose précisément d'étudier. Dès lors, les outils ADT peuvent être convoqués sur chaque sous-corpus : charge alors à l'analyste de mettre en comparaison des sorties-machines différentes mais de même facture (profils cooccurrentiels, rang de fréquences, AFC, etc.). Encore faut-il que les sous-corpus comparés soient comparables. A

défaut d'une langue commune et de locuteurs identiques, cette contribution a montré que le genre offrait cette comparabilité. Le genre transcende la langue et les auteurs. Il unifie ici notre corpus et rend analogues les prises de parole des présidents français ou brésiliens qui parlent au nom de leur pays. L'expérience ici réalisée sur le lexique pourrait être plus marquée encore sur d'autres observables linguistiques, sensibles aux genres, d'ordres grammaticaux ou syntaxiques.

Bibliographie

Benzécri J.-P. (1973). *L'Analyse des données*. T. 2 : *L'analyse des correspondances*. Dunod.

Brunet E. (2009). *Comptes d'auteurs. Études statistiques. De Rabelais à Gracq*. Champion.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. Larousse.

Halliday M. A. K. et Hasan R. (1976). *Cohesion in English*. Logman.

Iezzi S. *et al.* (editors) (2020). *Text analytics. Advances and Challenges*. Springer.

Lebart L., Pincemin B. et Poudat C. (2019). *Analyse des données textuelles*. P.U du Québec.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod

Mayaffre D. (2012). *Le discours présidentiel sous la Vème République*. Presses de Sciences Po.

Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles, in *JADT 2014*, Paris, 15-32.

Véronique J. (Ed.) (2000). *Parallel Text Processing: Alignment and use of translation corpora*. Kluwer Academic Publishers.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*. Champion.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique in *JADT 2000*, Lausanne, 171-178.

Online communication and political participation: a study dedicated to the referendum *Cannabis Legale*

Marina Marino, Rocco Mazza, Gabriele Luciano

University of Naples Federico II – m.marino@unina.it
rocco.mazza@unina.it gabriele.luciano97@gmail.com

Abstract
The digital transition undertaken by institutions invested political participation. Political participation refers to a range of activities that aspire to influence political decisions, the call for a popular referendum being one of them. The legally binding digital signature for requesting referendums and citizens' initiative laws was introduced in July 2021 with the approval of Law 77/2021. This decision established online as a concrete space for political participation. In this framework, social media covered a very important role, allowing the campaign to collect 500,000 signatures in only 7 days. The paper aims to analyze the issues of online political communication in relation to the socio-demographic characteristics of the subscribers of the first fully digital signature collection for the call for a referendum. We refer to the "Legal Cannabis" referendum. We applied an exploratory approach that used latent topic extraction methods to analyze social campaigns.

Keywords: Latent topic model, political participation, LDA

1. Introduction

In the Digital Era many aspects of life are integrated and supported by automatized and connected processes thanks to ICTs (Information and Communication Technologies). This trend is omnipresent: it involves even politics, from communication that has always been the cutting edge of technological transformation to new direct democracy forms. Our research study a fragmented phenomenon due to many research fields it could belong to and complex, the innovation processes transformed it in a political scenario. In September 2021 Italy experienced its first completely digital signatures collection to ask for a referendum, the campaign exceeded the threshold amount of 500.000 signatures with 607.627 signatures in less than two months. The request was a referendum on legalization of cannabis domestic cultivation, which is currently punished with jail from 6 to 20 years and an expensive fine. The goal of the work

is the exploration of communication aspects that emerged on social media within the collection time for the "Cannabis Legale" referendum campaign. Object of interest is the online debate and its contributors, to identify dynamics that move political initiatives involving direct democracy in the information society (Lyon and Ielasi, 1991). The concept of political participation represents the theoretical framework that defines this work. This identifies various activities that aim to influence policy decisions. They act in more or less direct ways and through legislative initiatives or peaceful or violent lobbying activities. The assumptions to support this focus are:

- The consolidation of digital signature for referendum proposals define the digital environment as a istitutional, political and participatory space. This transformation trained by the ICTs. The contemporary web technologies contributes at transitions, people as social media users can discover, promote, sign, debate, share with no need to do that outside the virtual world;
- Informations, news and political adhesion share the digital environment. There are also links between the online and offline practices. The political participation is an activity performed online and offline. In this way, social media represent public debate in the new public sphere constituted in the media system (Castells, 2008, 90).

In this study we assume that social media play a central role within the dynamics of political participation. From this point we will identify the main topics of discussion on the Twitter platform. The objective is to identify the main discussion topics that have animated the debate regarding the legal cannabis referendum.

2. Literature review

Social media are web-based technologies that encourage communicative exchange between users who use them. The first definition of these technologies dates back to Kaplan and Haenlein: "Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content. The concept of participation is very important to understand the role of these technologies within the major innovations of contemporary times. Grönlund defines participation as the act of doing things together or sharing events and narratives. Xie, Bo and Jaeger define political participation as "behaviours aimed at shaping governmental policy, either by influencing the selection of government personnel or by affecting their choices". The web is a collaborative medium and behaves more as a social medium than technological (Berners-Lee, 1999) and the entry of this tool for actions of democratic participation follows a process started long ago and that sees the boundaries between user and citizen

increasingly blur, allowing collective actions not possible before (Rheingold, 2002). Many studies show that the informational use of a wide range of media can be traced to modes of political participation. In this regard we refer to the use of newspapers (McLeod et al., 1999), television (Norris, 1996), the Internet (Shah, 2005), and mobile communication technologies (Campbell & Kwak, 2010). Another characteristic of online political participation is the timeliness with which it follows the public debate and the latest news, this evolution has only been possible with the advent of contemporary communication technologies, which have made information accessible in a timely manner (Best & Krueger, 2005). Online citizens can actively participate in political life and pursue their goals as active members of the community using the new tools made available by the online field. Such as sharing content with other users and sharing opinions on politics and current events, expressing dissatisfaction with governments by commenting on the social media pages of government officials, and participating in class actions against certain policies (Di Gennaro & Dutton, 2006). According to Gil de Zúñiga, Logan Molyneux, and Pei Zheng «the ease of use and creation of social media have generated an explosion of popular participation, allowing individuals to express their opinions more openly and freely as well as build a more active and meaningful relationship with official institutions».

3. Methodological strategy

The methodological procedure consists of the following steps: 1. Text collection and pre-processing; 2. Application of the model for topic extraction; 3. Content analysis. The texts were extracted using access to the API of the social network through a keywords based method. Starting from the corpus thus constituted, the pre-processing of the texts was performed following the following steps: 1. Normalization of the text; 2. Elimination of empty words (stopwords); 3. Elimination of symbols and non-alphabetic characters; 4. Extension of abbreviated expressions referring to norms, laws, codes and regulations. The texts that make up the corpus are public tweets shared by users over a defined range of time. During the normalization phase, characters were removed from the text through a unique encoding, punctuation marks, double spaces were removed and all characters were transformed into lower case. The reference dictionary was used for stopwords. At the end of the preprocessing phase, we created the document matrix. From this matrix, a model was applied to identify latent topics and reduce the information so that the main topics of discussion could be identified. Next, an interpetration was performed based on a critical reading of the model results by isolating the terms with higher probability for each topic. This reading was conducted using the conceptual references identified in our reference literature. Through the reduction of the information contained

in the corpus and the identification of latent topics, it was possible to explore and categorize the content of the texts under examination. Among the great variety of probabilistic topic model used to analyze the content of documents and the meaning of words (Blei et al., 2003; Griffiths and Steyvers, 2002; Hofmann, 2001), one worthy of attention is the Latent Dirichlet Allocation (from now LDA). LDA is a generative and unsupervised model, these characteristics were important for our choice to use it. We needed a method allows us to generate new texts, this texts represent the latent variables identified by the models. Our methodological objective is to study the semantic structure of this textual latent variables.

4. Data

We decided to explore the content produced by all users on the social network Twitter. For such phenomena of participation and activism, the actors are not only individual citizens. We also considered groups and associations that, on the basis of a shared collective identity, aggregate the public debate among heterogeneous actors who are related to each other (Diani, 1992). We selected five hashtags recognized among those most semantically contiguous to the subject matter of the referendum campaign: #referendumcannabis and #referendumcannabislegale, were officially used by the promoting committee and the organizations that were part of it, #cannabislegale is the hashtag that contains the message and the key terms of the referendum campaign, #megliolegale and #legalizziamo are instead two hashtags named after two civic promotion campaigns with the same name. The time frame considered for the analysis consists of the first 7 days of signature collection, between September 11 and 17, 2021. We collected 2778 texts. We decided to use this interval because the campaign reached the threshold of 500000 signatures. This threshold was defined based on the goals set by the promoters when they planned the campaign. It turns out to be an important goal reached in a very short time.

5. Political participation and social media, preliminary study

The political participation is the activity linked to aspiration to influence political decisions. Before conducting the textual analysis, the considerations made before on the relationship between political participation and social media have led us to a preliminary study dedicated precisely to this relationship. It is a multidimensional phenomenon, we operativize it in two ways: the signing of the referendum proposal, which empirically demonstrates support for the political initiative, and the publication of content on the subject by online users. We defined these two dimensions in accordance with the assumption that the two aspects (the signature and the publication of online content in support) are related. We decided a priori that the news of the signatures collected in a short period of

time was itself a reason to spread the initiative, making it reach new stakeholders causing a new increase in social media publications.

Table 1

Objective	Exploration of characterizing aspects of online communication for the campaign to collect signatures for the referendum *Legal Cannabis*.
Assumption	The web represents the space for public debate as a tool for social adherence.
Research question	What are the prevailing themes within the debate that has originated on social media?

The mutual influence between political participation and communication is outlined in the relationship between digital subscription and information disseminated online: the assumption is that the more the collection of signatures proceeds, the more the information spreads online, and at the same time, as the circulation of information increases, the number of subscriptions increases. We studied the correlation between the number of online contents and the signatures recorded in the time range under analysis. We used the number of contents in the downloaded data stream and the number of signatories to the proposal. These trends have a strong linear correlation (Figure 1). The Pearson correlation coefficient $\rho = 0.964$ indicates that the correlation is extremely positive.

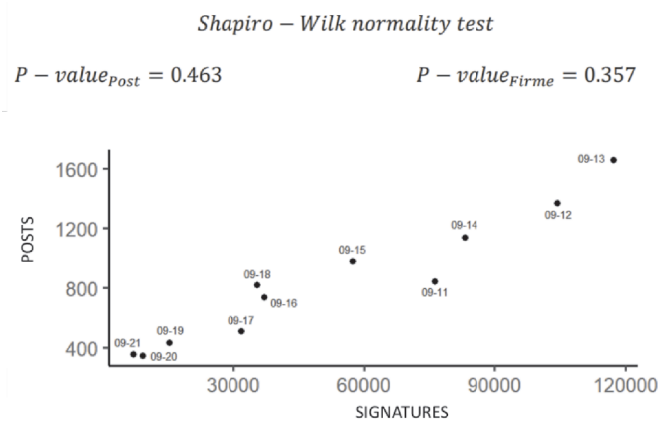


Figure 1. Relation between the number of posts and digital signatures

The interpretive framework is based on the theoretical model recognized by Cernison (2019, 34) by investigating the world of online activism as an environment in which the practices of participation are contaminated to the typical communication practices of the considered medium. Moreover, Cernison

(ibid.) adds that online communication is not only considered as an element that can influence society, but also as a complex environment of communicative action in which organizations adopt social media platforms as tools of political action. These preliminary considerations allowed us to proceed from an social media participation study perspective, in which the results that emerged from the online content analysis explained the themes of public debate about the referendum.

6. Results

To determine the number of arguments (K) we provides four metrics by computing according to the algorithms of Griffiths (2004), CaoJuan (2009), Arun (2010) and Deveaud (2014). We selected K = 5.

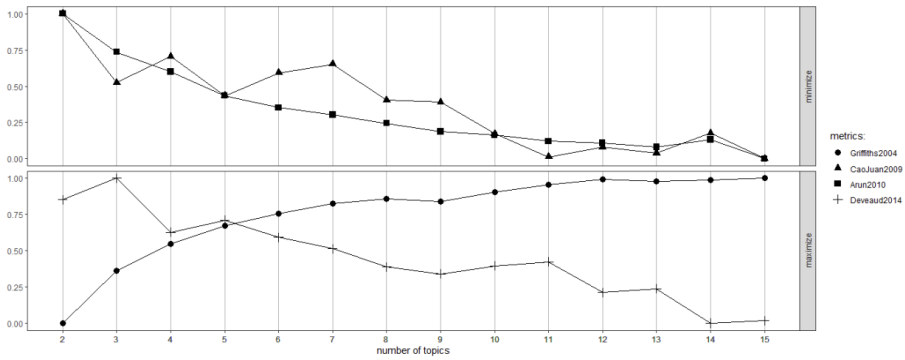


Figure 2. Topic number calculation

The analysis extracted the prevailing themes within the processed content (Tab. 2). We isolated a first topic that represents demonstrations by users of support for the initiative; this topic includes posts published by campaign supporters not actively involved within the committees. So this topic collects the messages of those who participated in the debate expressing their ideas exclusively online. The word politics invites us to consider the central role of political institutions in the debate in support of legalization. The second topic is dedicated to the fight by the state against the mafias that control the distribution of cannabis under a strictly punitive regulation. This topic is an environment of confrontation in which we discuss the positive consequences of legalization in the fight against mafias. The third topic describes the progress of the campaign and how the collection of signatures is proceeding, with reference to the milestones reached. Very important words such as legalization and mafias represent the main themes of the referendum campaign. The fourth topic provides news about the campaign

with references to how to sign. The fifth topic is dedicated to anti-prohibitionism and the importance of carrying out the demands of the campaign. The link to the signature portal is very important and creates a link between the social and the act of signing.

Table 2 – Most associated words for each topic

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
politica	legale	firmato	firme	sito
settembre	stato	firma	firmare	storia
roma	firmato	firmare	cannabis	link
bene	firme	online	referendum	legalizzala
dover	droga	giorni	spid	legge
visto	posizione	tweet	firmato	segretario
persone	persone	parlato	firma	civilta
dire	capire	raggiunt	mafie	partito
vero	pensar	tendenza	raccolta	onorevole

7. Further developments

The work presented represents part of a preliminary study. We will extend the research by including within the design data collected from other social networks, such as Instagram. A further future development of the work will be to integrate the data from social with the master data of the signatories, to provide a complete and integrated picture of the phenomenon studied.

References

Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 59–68

Grönlund Å. (2009). ICT Is Not Participation Is Not Democracy: eParticipation Development Models Revisited. In: Macintosh, A., Tambouris, E. (eds.) ePart 2009. LNCS, vol. 5694, pp. 12–23. Springer, Heidelberg

Xie B. and Jaeger P.T. (2008). Older Adults and Political Participation on the Internet: A Crosscultural Comparison of the USA and China. Journal of Cross Cultural Gerontologyurnal of Cross Cultural Gerontology, 1–15

McLeod, J. M., Scheufele D. A., Moy P., Horowitz E. M., Holbert R. L., Zhang W. and Zubric J. (1999). Understanding deliberation the effects of discussion networks on participation in a public forum. Communication Research, 26(6), 743–774.

Norris P. (1996). Does television erode social capital? A reply to Putnam. Political Science and Politics, 29(3), 474–480. Retrieved from http://www.jstor.org/stable/420827

Shah D. V., Cho J., Eveland W. P. and Kwak N. (2005). Information and expression in a digital age modeling Internet effects on civic participation. Communication Research, 32(5), 531–565. doi:10.1177/0093650205279209.
Campbell S. W. and Kwak N. (2010). Mobile communication and civic life: Linking patterns of use to civic and political engagement. Journal of Communication, 60(3), 536–555. doi:10.1111/j.1460-2466.2010.01496.x.
Best S. J. and Krueger B. S. (2005). The effect of Internet use on political participation: An analysis of survey results for 16-year-olds in Belgium. Social Science Computer Review Winter, 26, 411–427. doi:10.1177/0894439307312631.
Di Gennaro C. and Dutton W. (2006). The Internet and the public: Online and offline participation in the United Kingdom. Parliamentary Affairs, 59(2), 299–313. doi:10.1093/pa/gsl004.
Gil de Zúñiga H., Molyneux L. and Zheng P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. Journal of communication, 64(4), 612-634.
Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation, Journal of machine Learning research, 3(Jan), 993-1022.
Hofmann T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2), 177-196.
Griffiths T. L. and Steyvers M. (2002). A probabilistic approach to semantic representation. In: Proceedings of the annual meeting of the cognitive science society, Vol. 24, No. 24.
Lyon D. and Ielasi P. (1991). La società dell'informazione. Il Mulino.
Castells M. (2008) The New Public Sphere: Global Civil Society, Communication Networks, and Global Governance. The ANNALS of the American Academy of Political and Social Science, 616(1), 78–93.

Clustering of financial time series: a bibliometric analysis

Raffaele Mattera[1], Michelangelo Misuraca[2],
Germana Scepi[3], Maria Spano[3]

[1] Sapienza University of Rome – raffaele.mattera@uniroma1.it
[2] University of Calabria – michelangelo.misuraca@unical.it
[3] University of Naples "Federico II"– scepi@unina.it
maria.spano@unina.it

Abstract
Clustering is nowadays widely applied in finance (Mantegna (1999), for solving portfolio selection and risk management problems. In this paper, we propose a review related to both state-of-the-art and of the recent developments of this approach. We adopt a bibliometric analysis, mapping the main issues discussed by scholars in the last 30 years with a network-based technique known as thematic analysis.

Keywords: Clustering, Machine learning, Unsupervised learning, Financial time series, Thematic Analysis

1. Introduction

Nowadays, financial markets are considered complex systems characterised by many components that interact continuously. This peculiarity makes them very difficult to modelling. Furthermore, the presence of stylised facts (see Cont, 2001) in many financial time series causes a lot of problems in their dealing with. The existence of a peculiar clustering structure (Mantegna, 1999) in financial markets has important theoretical and empirical implications (e.g. Ando and Bai, 2017; Barak and Modarres, 2015; Zhong and Enke, 2017; Duarte and Castro, 2020; Raffinot, 2017; Iorio et al., 2018) and several studies have been developed on clustering approaches for financial data. The different papers share some elements. The two most frequently discussed aspects consist in the definition of a proper distance among the units (Fan et al., 2016) and in the algorithm's selection. These two main points are often dealt with together.
Concerning the choice of an appropriate dissimilarity measure, the proposed clustering methods can be categorised into two main classes: feature-based and model-based approaches (Maharaj et al., 2019). The main advantages of feature-

based clustering approaches are the dimensional reduction generated and their useful application to time series of different lengths. In the model-based approaches, the time series are assumed to be generated from specific underlying models, or by a combination of probability distributions, and the similarity between fitted models is evaluated.

Dealing with financial time series, in the class of feature-based clustering approaches, Mantegna (1999) proposed the use of the correlation coefficient as similarity measure but, Caiado et al. (2006) and D'Urso and Mahraj (2009) observed that it does not consider the dynamic structure of the time series data. These authors preferred using the Auto-Correlation Function (ACF) and the periodogram ordinates, in the frequency domain. Alternative feature-based distances based on cepstral (D'Urso et al., 2020) or wavelet coefficients (Maharaj et al., 2010) have been proposed.

In the framework of financial data, different studies have shown as the use of model-based approaches has the advantage of capturing the "asymmetric volatility" of the data. Therefore, several distances have been proposed, such as the GARCH-based distances (D'Urso et al., 2013, 2016; Otranto, 2008), the distances based on multivariate dynamic covariance models (e.g. DCC as in Otranto, 2010) and the MEM-based distances (Otranto and Gargano, 2015; Gallo et al., 2021). Another important model-based distance considers the distribution of stock returns time series. De Luca and Zuccolotto (2011, 2021) analysed the dissimilarities of the stock returns by looking at the tails' behaviour under a specified copula. More recently, Cerqueti et al. (2021) exploited the time variation in the distribution of higher moments to define financial time series dissimilarity. Other proposed approaches are based on functional representation, such as spline parameters (Cairoli et al., 2005; Iorio et al., 2018).

Looking at the clustering algorithm choice for financial data (see Fan et al., 2016 for a summary), two main ways have been adopted: the use of hierarchical-based clustering algorithms (Caiado and Crato, 2010; De Luca and Zuccolotto, 2011, 2021; Mantegna, 1999; Otranto, 2008) and the use of partitioning based clustering ones, such as k-means, k-medoids and their fuzzy versions (Cerqueti et al., 2021; D'Urso, 2016, 2020; Iorio et al., 2018). However, also the model-based approach developed by Fraley and Raftery (2002) and Fröhwirth-Schnatter and Kaufmann (2008) has been applied to financial time series (e.g. Wang and Tsay, 2019) as well as the Genetic Algorithm (e.g. Pattarin et al., 2004) and Self Organizing Maps (SOM, see Silva, and Marques, 2010).

This paper provides a first bibliometric analysis of the clustering financial time series topic. The rest of the paper is structured as follows. In Section 2, the data and the applied methodology are shown. The main results are discussed in Section 3.

2. Data and methodology

Bibliometrics allows investigating the written scientific production of a research domain by introducing a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity (Broadus, 1987; Diodato, 1994). It involves quantitative methods for exploring, monitoring, and measuring scientific research. Many research fields use bibliometric methods to explore the impact of their field, the impact of a set of researchers, the impact of a particular paper, journals taken as a reference by researchers, the input knowledge, research gaps, trends, and future opportunities (Tsay, 2013; Zaho, 2010).

The key procedures commonly implemented in bibliometric studies are performance analysis (Peters and van Raan, 1991; White and McCain, 1998) and science mapping (Börner et al., 2003; Noyons et al., 1999). The first one aims to evaluate the productivity and the popularity of the different actors (e.g. authors, institutions, countries) based on bibliographic data. The second one, science mapping, allows presenting a "big picture" of existing research on the investigated domain, highlighting its main topics and their evolution from a diachronic perspective (Cobo et al., 2011; Garfield, 1994). Mapping techniques frequently refer to textual data analysis. Each topic can be characterised through the set of keywords assigned by the publications' authors or by the citation indexes. Starting from this keyword set, it is possible to build a co-occurrence matrix to represent the collection of publications. Network tools can be then used to visualise and analyse the knowledge base (Aria et al., 2020).

To collect the publications about time series clustering over the past 30 years, on February 18th, 2022, we queried the Web of Science (WoS) indexing database - now maintained by Clarivate Analytics. We searched by using the following query "*Cluster** ***OR*** *Hierarchical Structure**" **AND** "*Financial time series* ***OR*** *stock market**" ***NOT*** "*Volatility Clustering*", obtaining 698 entries in the period 1992–2021. It is important to say, however, that WoS indices only the publications that appeared after 1985. We selected original articles, review articles, book chapters and proceedings published in English. The analysis was performed by using the R library *bibliometrix* (Aria and Cuccurullo, 2017).

In Figure 1, the main descriptive statistics about the analysed collection are reported. In the reference interval 1992-2021, a total of 698 documents have been published by 1598 authors. A total of 80 documents have been written by a single author, while 618 are multi-authored documents. The documents received a total of 19536 citations, leading to approximately 21 citations per publication.

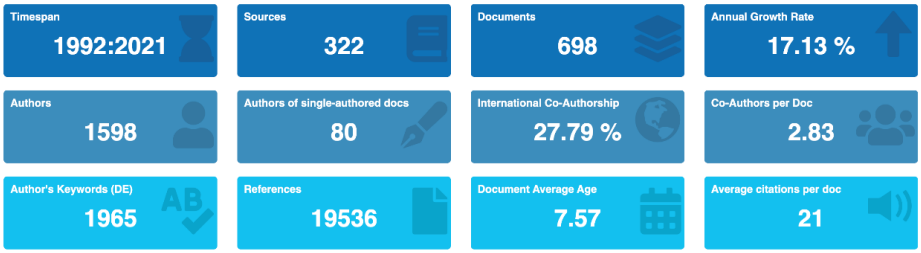


Figure 1. Main statistics about the analysed collection

3. Main Results

Considering the thematic map in the first selected period (1992-2021), we find the presence of very few topics. This is an expected result since the financial time series clustering literature grew with very low rates up to 2002 (see Fig. 2). From 2003, we observe a particular rapid increase in scientific production.

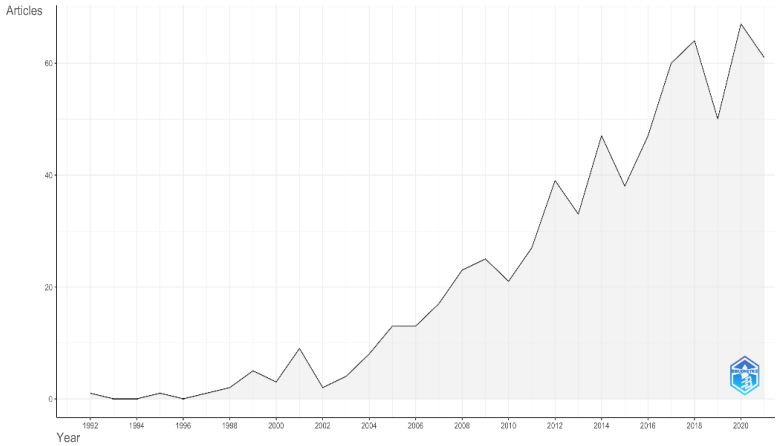


Figure 2. Annual scientific production between 1992 and 2021

The thematic maps related to the second time slice, 2003-2012 (see Fig. 3), show that "econophysics" and "clustering" are basic themes, suggesting that during this period most of the studies devoted to clustering of financial time series have been published by econophysicists.

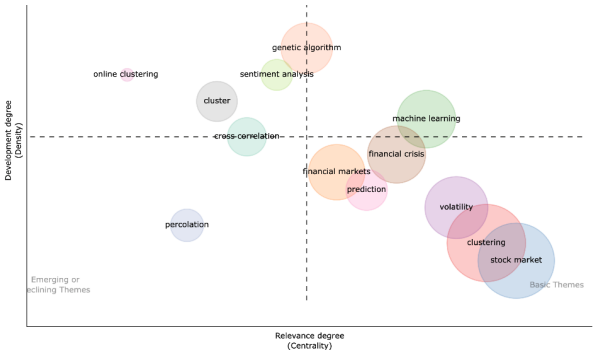


Figure 3. Thematic map during the years 2003-2012

The k-means algorithm represents an important motor theme during this period, suggesting the relevance of partitioning approaches in clustering financial time series. Other two important motor themes can be identified by the market microstructure and cross-correlation, showing the direction of the literature in the definition of the time series dissimilarities.

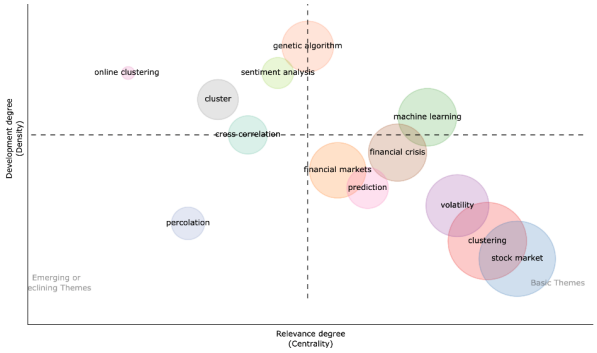


Figure 4. Thematic map during the years 2012-2021

Further, behavioural finance has become an important emerging theme in terms of applying clustering techniques. The thematic map related to the last period 2013-2021 (see Fig. 4), is characterised by the absence of the "econophysics" theme. This result can be explained by the authors of statistics and economics that published a lot of articles on this topic. The motor theme in the last period is the clustering of financial time series in terms of volatility patterns (i.e. "clustering", "volatility", and "stock market" clusters). The clustering approach based on cross-correlation becomes a declining topic while an important new topic "machine learning" becomes a motor one.
The last map summarises the new challenges of financial time series clustering,

given by the development of clustering approaches for financial time series in data-rich environments. An example is given by online clustering approach representing a niche topic in thematic maps but with the potentiality to become a motor, in particular for real-time portfolio selection tasks.

4. Final remarks

The identification of common structures and patterns in financial market is a challenging task. Previous studies developed different clustering techniques with the aims of highlighting the presence of a common structures in the financial market. Nowadays, cluster analysis is also used to improve portfolio selection performances. In this paper we propose the first literature review on financial time series clustering with the use of a quantitative technique. We find that there is an increasing interest on this topic and that the use of machine learning algorithms for clustering financial time series is an emerging topic in the last years.

References

Ando, T., and Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519): 1182-1198

Aria, M., and Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis, *Journal of Informetrics*, 11(4): 959-975

Aria M., Misuraca M., and Spano M. (2020). Mapping the evolution of social research and data science on 30 years of Social Indicators Research, *Social Indicators Research*, 149: 803–831

Bastos, J. A., and Caiado, J. (2021). On the classification of financial data with domain agnostic features. *International Journal of Approximate Reasoning*, 138: 1-11

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications*, 42(3):, 1325-1339

Caiado, J., Crato, N., and Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10): 2668-2684.

Caiado, J., and Crato, N. (2010). Identifying common dynamic features in stock returns. *Quantitative Finance*, 10(7): 797-807

Cerqueti, R., Giacalone, M., and Mattera, R. (2021). Model-based fuzzy time series clustering of conditional higher moments. *International Journal of Approximate Reasoning*, 134: 34-52.

Cerioli, A., Laurini, F., and Corbellini, A. (2005). Functional cluster analysis of financial time series. In *New Developments in Classification and Data Analysis*, pp. 333-341

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2): 223

D'Urso, P., and Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24): 3565-3589

D'Urso, P., De Giovanni, L., Massari, R., D'Ecclesia, R. L., and Maharaj, E. A. (2020). Cepstral-based clustering of financial time series. *Expert Systems with Applications*, 161:

113705
De Luca, G., and Zuccolotto, P. (2011). A tail dependence-based dissimilarity measure for financial time series clustering. Advances in Data Analysis and Classification, 5(4), 323-340
De Luca, G., and Zuccolotto, P. (2021). Hierarchical time series clustering on tail dependence with linkage based on a multivariate copula approach. *International Journal of Approximate Reasoning*, 139: 88-103
Duarte, F. G., and De Castro, L. N. (2020). A framework to perform asset allocation based on partitional clustering. *IEEE Access*, 8: 110775-110788
Fan Cai, Nhien-An Le-Khac and M-Tahar Kechadi (2016). Clustering Approaches for Financial DataAnalysis: a Survey, *arXiv*: 1609.08520
Fraley, C., and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458): 611-631
Fröhwirth-Schnatter, S., and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1): 78-89
Gallo, G. M., Lacava, D., and Otranto, E. (2021). On classifying the effects of policy announcements on volatility. *International Journal of Approximate Reasoning*, 134: 23-33
Iorio, C., Frasso, G., D'Ambrosio, A., and Siciliano, R. (2018). A P-spline based clustering approach for portfolio selection. *Expert Systems with Applications*, 95: 88-103
Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern recognition*, 38(11): 1857-1874
Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1): 193-197
Maharaj, A. E., D'Urso, P., and Galagedera, D. U. (2010). Wavelet-based fuzzy clustering of time series. *Journal of Classification*, 27(2): 231-275
Maharaj, E. A., D'Urso, P., and Caiado, J. (2019). *Time series clustering and classification*. Chapman and Hall/CRC
Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, 52(10): 4685-4698
Otranto, E. (2010). Identifying financial time series with similar dynamic conditional correlation. *Computational Statistics & Data Analysis*, 54(1): 1-15
Otranto, E., and Gargano, R. (2015). Financial clustering in presence of dominant markets. *Advances in Data Analysis and Classification*, 9(3): 315-339
Pattarin, F., Paterlini, S., and Minerva, T. (2004). Clustering financial time series: an application to mutual funds style analysis. *Computational Statistics & Data Analysis*, 47(2): 353-372
Raffinot, T. (2017). Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management*, 44(2): 89-99
Silva, B., and Marques, N. C. (2010). Feature Clustering with Self-organizing Maps and an Application to Financial Time-series for Portfolio Selection. In *IJCCI (ICFC-ICNC)*, pp. 301-309
Zhong, X., and Enke, D. (2017). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, 267: 152-168
Wang, Y., and Tsay, R. S. (2019). Clustering multiple time series with structural breaks. *Journal of Time Series Analysis*, 40(2): 182-202

Contribution à l'histoire d'un concept : occcurrences et cooccurrences de 'souveraineté' dans les discours de Charles de Gaulle (1958-1969) et d'Emmanuel Macron (2017-2022)

Damon Mayaffre

Université Côte d'Azur – damon.mayaffre@univ-cotedazur.fr

Abstract
In France, 'sovereignty' has been a key concept in political thought since Jean Bodin's *Six Books of the Republic* (1576). This contribution aims to contribute to the recent history of the concept, based not on the dictionary but on a large corpus of contemporary discourse (1958-2022). On the one hand, the study of the statistical distribution of the term attests to an uneven and atypical use over the 70 years of the corpus. On the other hand, the study of co-occurrences shows a major evolution of the co-(n)text of use from President de Gaulle's speeches to President Macron's speeches. Finally, the return to the text shows how qualifying adjectives (e.g. 'digital sovereignty') succeed in the 2020s the initial noun complements of the 1950s (e.g. 'the sovereignty of the people'). To the point of changing the meaning of the concept?

Keywords: logometry, textometry, lexicometry, sovereignty, co-occurrences, specificities, concordances.

Résumé
La souveraineté est, en France, depuis *Les Six Livres de la République* de Jean Bodin (1576) un concept clef de la pensée politique. Cette contribution entend participer à l'histoire récente du concept, en s'appuyant non pas sur le dictionnaire mais sur un gros corpus de discours contemporain (1958-2022). D'une part, l'étude de la distribution statistique du terme atteste d'un usage inégal et atypique au cours des 70 ans du corpus. D'autre part, l'étude des cooccurrences atteste d'une évolution majeure du co(n)texte d'utilisation des discours du président de Gaulle aux discours du président Macron. Enfin le retour au texte montre comment les adjectifs qualificatifs (par exemple 'souveraineté numérique') succèdent dans les années 2020 aux compléments du nom initiaux des années 1950 (par exemple 'la souveraineté du peuple'). Jusqu'à changer le sens du concept ?

Mots clés : logométrie, textométrie, lexicométrie, souveraineté, cooccurrences, spécificités, concordances

1. Introduction

Depuis les travaux de lexicométrie politique du Laboratoire de Saint-Cloud, nous savons que le débat politique se gagne ou se perd d'abord par le choix et la répétition des mots ; choix que la statistique occurrentielle met bien en valeur. Ensuite, et plus encore peut-être, le débat politique se gagne ou se perd par le sens que le locuteur politique réussit à imposer aux mots ci-devant répétés ; sens que la statistique cooccurrentielle permet d'approcher. Imposer ses mots notamment par l'effet de répétition, leur prescrire un sens adéquat par effet de co(n)textualisation : voilà le double enjeu de l'orateur politique (et à sa suite de l'analyste) lorsqu'il exprime son programme et son idéologie (pour la littérature saint-clousienne par ex. Tournier, 1997).

Dans le cadre de l'histoire des concepts de (Skinner, 1978), (Kosseleck, 1990) ou (Guilhaumou, 2006), cette contribution étudie dans le corpus présidentiel français sous la V^{ème} République (1958-2022 – 1000 discours – 4,5 millions de mots), la notion de 'souveraineté' ; notion historique et centrale en République depuis Jean Bodin, Jean-Jacques Rousseau ou Emmanuel-Joseph Sieyès ; notion toujours cardinale aujourd'hui dans la Constitution française actuelle qui y consacre ses premiers articles et son Titre premier : « De la souveraineté »[1].

Le traditionnel *calcul des spécificités* d'une part et les graphes de cooccurrences implémentés dans le logiciel *Hyperbase* (Brunet, 2012 ; Vanni, 2016 d'autre part permettent de dire quand et comment on parle de la 'souveraineté' dans notre recueil de textes. La navigation hypertextuelle dans le corpus et les concordanciers, ainsi qu'une approche grammaticale au fil des mots, permettent de retourner au texte et à la phrase pour une lecture augmentée et un sens, ou plutôt *des* sens, non pas formalisés en dictionnaire mais représentés en corpus.

2. 'Souveraineté' : la distribution inégale d'un concept

Dans le corpus présidentiel français (1958-2022), rares sont les mots qui se distribuent quantitativement comme le mot 'souveraineté'. Habituellement, comme le montrent les travaux de (Salem, 1991) sur les corpus chronologiques, le temps, c'est-à-dire l'actualité et l'histoire, impose le rythme d'apparition du lexique : les mots, *grosso modo*, croissent ou déclinent de manière régulière et continue ou, à strictement parler, de manière *chrono-logique*.

Or ici, tout au contraire, et selon le *calcul des spécificités* (Lafon, 1984), le mot est sur-utilisé par de Gaulle au tout début de la période puis par Macron à 60 ans de distance à l'autre extrémité de la période. Entre ces deux moments, le terme

[1] Pour le droit français actuel, on consultera l'ouvrage de référence : Arnaud Haquet, Le concept de souveraineté dans le droit constitutionnel français, PUF, 2006.

est sous-utilisé aussi bien par Pompidou que Giscard, Mitterrand que Chirac, Sarkozy que Hollande (figure 1).

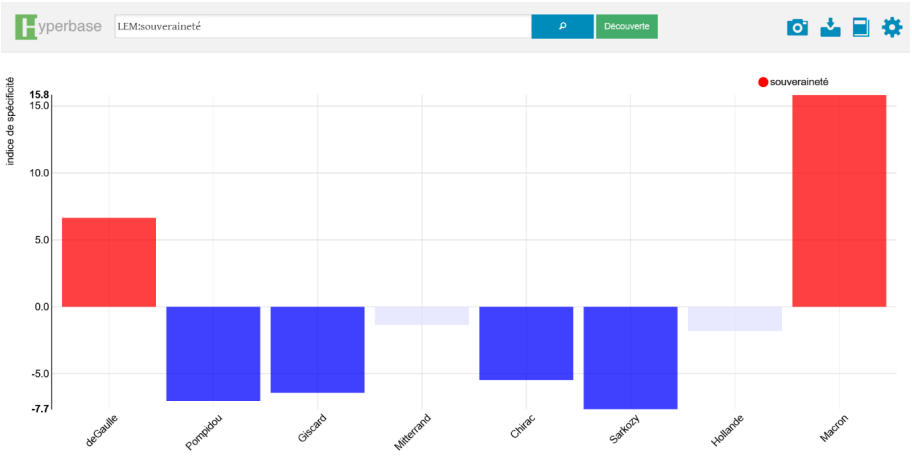


Figure 1. Distribution du lemme 'souveraineté' dans le corpus présidentiel (indices de spécificité exprimé en écarts réduits) – Hyperbase 2022.

Ainsi pouvons-nous faire l'hypothèse que l'époque contemporaine connait, en France, deux moments essentiels autour de la question de la 'souveraineté' : la naissance de la V^{ème} République jusqu'à mai 1968 d'une part, durant laquelle se discute la légitimité démocratique d'un nouveau régime présidentiel, puis, un demi-siècle après, un deuxième moment, durant lequel la question démocratique semble se poser à nouveau, peut-être à propos du mouvement populaire des Gilets jaunes et de la défiance des citoyens vis-à-vis du politique (abstention massive, désaffiliation partisane, désyndicalisation, etc.), peut-être à propos de la construction européenne qui interroge le fait politique national.
Mais face à ce constat occurrentiel, faut-il dire que de Gaulle et Macron ont, en dépit du temps qui s'écoule, un discours identique ? Derrière la remobilisation d'un signifiant doit-on supposer une remobilisation du signifié ? Assistons-nous simplement à un bégayement lexical de l'histoire sur un concept qui reviendrait régulièrement dans le discours et serait un inchangé de la pensée politique ?

3. Le sens évolutif d'un concept

Pour répondre à ces questions, nous convoquons les cooccurrences de 'souveraineté' respectivement chez de Gaulle (1958-1969) et chez Macron (2017-2022).
En s'appuyant sur le grand courant anglo-saxon contextualiste de (Firth, 1957), de (Halliday et Hasan, 1976) ou de (Sinclair, 1991) nous essayons d'implémenter le célèbre : « you shall know a word by the company it keeps ». Ainsi, c'est non plus comme une simple pétition de principe qualitative mais comme une

implémentation effective, quantitative et statistique, que nous avons posé aux JADTs de Paris en 2014 que *le sens d'un mot est la somme de ses cooccurrences* (Mayaffre, 2014) ; et que l'unité de sens de la lexicométrie était non pas l'item ou le token mais la paire cooccurrentielle (Mayaffre, 2008).
Un nuage de mots des *cooccurrents spécifiques* (comme précédemment : Lafon, 1984)[2] de 'souveraineté' laisse comprendre l'ampleur de l'évolution conceptuelle en cours ; précisons simplement que nous avons filtré les résultats sur les seuls noms et adjectifs (figure 2)

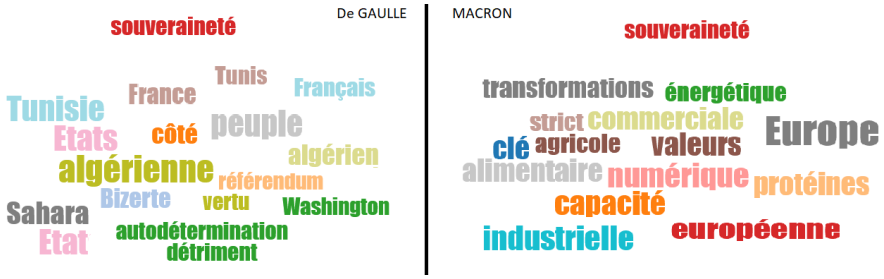


Figure 2. Cooccurrences spécifiques de 'souveraineté' chez de Gaulle (à gauche) et Macron (à droite). La taille des mots est proportionnelle à l'indice de cooccurrence ; la couleur des mots est seulement esthétique. (Hyperbase 2022)

'Souveraineté' se contextualise (ie. cooccurre statistiquement) avec des mots systématiquement différents chez de Gaulle et chez Macron. Sa charge ou coloration sémantique s'en trouve modifiée. Les univers lexicaux que 'souveraineté' polarise ou dans lesquels 'souveraineté' gravite divergent fondamentalement. Chez de Gaulle, 'souveraineté' renvoie systématiquement au paradigme de la République et de l'Etat-Nation dans le droit fil de l'historique définition de Jean Bodin dans *Les Six Livres de la République* (1576) ; cet Etat-Nation est bien sûr chez de Gaulle celui de la France ('France', 'français') mais aussi, dans une période de tensions décoloniales, celui de l'Etat-Nation algérien, tunisien, marocain ('algérien', 'Tunis', 'Sahara').
Chez Macron, en revanche, 'souveraineté' renvoie à une pluralité de paradigmes. L'Europe a remplacé la France ('Europe', 'européenne') mais surtout d'autres

[2] Nous savons les propositions riches, fortes et nombreuses à l'échelle internationale pour calculer les phénomènes coocurrentiels (indice de Church, indice de Dunning, indice de Brunet, etc.), et savons la variété des représentations visuelles de ces indices (listes, tableaux, arbres, graphes, etc.) Ici le calcul des cooccurrents spécifiques et la représentation dans un nuage où la taille des mots est proportionnelle à l'indice apparaissent simples et efficaces pour notre démonstration.

champs non nécessairement liés à la République ou à l'Etat-nation s'imposent ('alimentaire', 'numérique', 'commerciale', 'industrielle', 'agricole', etc.). En un mot, de Gaulle inscrit le concept dans une réflexion régalienne et géo-politique, là où Macron diversifie le point de vue et accorde à l'économique une place prépondérante.

4. De l'autorité absolue au concept qualifié

Enfin le retour au texte, via le concordancier, prolonge le parcours de lecture (figure 3 et 3-bis) et, à consulter toutes les occurrences de 'souveraineté' en contexte (KWIC - Pincemin, 2006) l'impression est saisissante (figure 3 et figure 3-bis)

France , qui ne doive revendiquer sans relâche la	souveraineté	algérienne sur le Sahara . Enfin le fait que
, il y a l' avènement à la	souveraineté	d' un grand nombre d' Etats qui se sont
prises parce qu' ils n' étaient pas sous la	souveraineté	de la France , ils étaient sous sa tutelle
nous pouvions penser que , tout en proclamant la	souveraineté	de la Tunisie sur Bizerte , souveraineté qui n'
argument le plan français , qui maintient la	souveraineté	des Etats , ne répond pas à notre conception
et qui , faute des ressorts que sont la	souveraineté	des peuples et la responsabilité des Etats , serait
rre étrangère apparut la République ! Elle était la	souveraineté	du peuple , l' appel de la liberté ,
en œuvre , quand il le faut , la	souveraineté	du peuple , le président détient en permanence la
et notre Constitution , laquelle fixe que " la	souveraineté	française appartient au peuple français , qui l' exerce
. C' est la plus justifiée , car la	souveraineté	nationale appartient au peuple et elle lui appartient évi

Figure 3. Concordances de 'souveraineté' chez de Gaulle (extraits)

Chez de Gaulle la 'souveraineté' semble presque toujours suivie d'une préposition (*d', du, de*) : le nom a ainsi son complément ('souveraineté du peuple', 'souveraineté de la France', 'souveraineté des Etats', etc.). Chez Macron, la 'souveraineté' est le plus souvent suivie d'un adjectif : le nom est qualifié ('souveraineté numérique', 'souveraineté alimentaire', 'souveraineté européenne', etc.). Statistiquement, sur notre corpus étiqueté morphosyntaxiquement, l'impression pourrait se chiffrer avec les distributions effectivement inégales du motif 'souveraineté + PREPOSITION' (plus fréquent chez de Gaulle) *versus* 'souveraineté + ADJECTIF' (plus fréquent chez Macron). Du complément de nom chez de Gaulle à l'adjectif chez Macron, l'évolution linguistique doit être travaillée.

. Je pense toujours que nous devons rebâtir une	souveraineté	industrielle , numérique et agricole et pour cela investir
pris vis-à-vis de nos agriculteurs . Je veux une	souveraineté	alimentaire européenne . Nos agriculteurs ont raison de pas
soit un plan de relance industriel pour bâtir notre	souveraineté	industrielle et donc faire relocaliser certaines industries ; éc
aidant , en investissant , on va bâtir une	souveraineté	agricole qui n' existe pas . Aujourd'hui , j'
précisément poursuivre notre travail pour plus de	souveraineté	numérique , industrielle , économique en France et en
nos émissions de Co2 , compatible avec notre	souveraineté	industrielle et technologique . Ce nouveau modèle
de notre modèle , de nos valeurs . Cette	souveraineté	culturelle , académique , industrielle , de recherche ,
c' est ça qui est pertinent en termes de	souveraineté	économique et en termes d' émissions de CO2 .
un ordre républicain juste , la défense de la	souveraineté	européenne . Oui , il faut toujours qu' il
cadre , et nous devons construire , une nouvelle	souveraineté	européenne par laquelle nous apporterons la réponse claire

Figure 3-BIS. Concordances de 'souveraineté' chez Macron

Chez de Gaulle la 'souveraineté' est *attribuée*. Elle apparait ainsi avant tout comme un *exercice* au sens fort ou une prérogative ; la prérogative d'êtres animés (les Français, le peuple) ou d'incarnation (la 'France' dans le cadre d'une prosopopée, 'Washington' dans le cadre d'une antonomase). La souveraineté est donc un pouvoir qui s'exerce, particulièrement le pouvoir de décision du peuple. Chez Macron, la souverainété n'est pas attribuée mais est *qualifiée*. C'est moins son exercice qui compte que son domaine d'application. La souveraineté se comprend et se nuance par son champ d'application.

En convoquant 'souveraineté' telle qu'elle est, de Gaulle semble s'appuyer sur un allant de soi conceptuel et historique stabilisé depuis Jean Bodin. En qualifiant le nom par divers adjectifs, Macron, lui, décline l'idée, mais prend le risque de l'affaiblir.

On prête en effet à (Barthes, 1953) dans le *Degré zéro de l'écriture*, l'idée que l'adjectif est la mort du concept – ainsi fait-on fondre le concept de 'neige' si on croit nécessaire de préciser 'neige blanche' ou 'neige froide'. Admettons en tout cas que le besoin de qualifier 'souveraineté' par divers adjectifs revient à la doter d'attributs spécifiques mais prend le risque de diluer ses attributs ontologiques. Le caractère absolu de la 'souveraineté' que l'on retrouve dans les discours de Gaulle laisse ainsi la place à des sens plus concrets, particuliers, discrétionnaires. Le concept et son empan sémantiques - disons sa définition - ne vont plus de soi. Jusqu'à prétendre, au fond, que le concept se déconceptualise ?

5. Conclusion

L'histoire des concepts en Europe est un plaidoyer en faveur du corpus et de l'archive. Le sens que l'on croit stabilisé, en synchronie, dans les dictionnaires d'une époque, se réalise en réalité par la force de l'usage en discours, par les combats politiques menés, par la répétition et la co(n)textualisation.

Le travail sur la cooccurrence permet d'objectiver cette sémantique de corpus, par définition variable en fonction de l'époque et du locuteur. Par simple calcul

et représentation graphique, il est possible de constater le halo sémantique qui entoure les termes : halo duquel se dinstingue finalement le sens. Resterait alors la nécessité de poursuivre la démarche, avec des approches cooccurrentielles toujours plus ambitieuses sur les profils cooccurrentiels ou les cooccurrents de deuxième ou de troisième ordre (Bertels et Speelman, 2012; Vanni, 2016). Par exemple, maintenant que l'on a admis que chez de Gaulle la 'souveraineté' est l'exercice du peuple, la question des modalités de cet exercice devient essentielle. Le graphique des cooccurrents de second ordre, élargi à l'ensemble des mots (et pas seulement), autour du mot pôle 'souveraineté', puis de son cooccurrent principal 'peuple', nous met sur la voie d'un discours républicain mais bonnapartiste, dans lequel la 'souveraineté' 'appartient' au 'peuple' et 's'exerce' avant tout par l'usage du 'referendum' (figure 4)

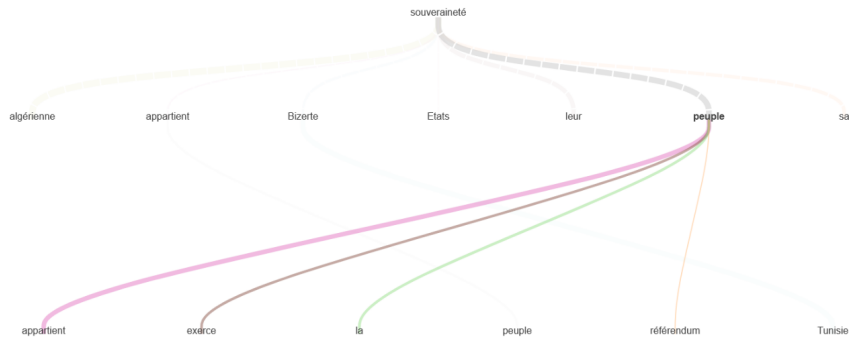


Figure 4. Graphe de cooccurrences de 1er et 2ème ordre (l'épaisseur des traits est proportionnelle à l'indice de cooccurrence). Hyperbase 2022.

Bibliographie

Barthes R. (1953). *Degré zero de l'écriture*, Seuil.
Bertels A. et Speelman D., La contribution des cooccurrences de deuxième ordre à l'analyse sémantique, *Corpus* (En ligne), 11.
Brunet E. (2012), Nouveau traitement des cooccurrences dans Hyperbase, *Corpus* (En ligne), 11
Firth R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.
Guihaumou J. (2006), *Discours et événement. L'histoire langagière des concepts*. Presse universitaire de Franche-Comté.
Haquet A. (2004). *Le concept de souveraineté en droit constitutionnel français*. Puf.
Halliday M. A. K. et Hasan R. (1976), *Cohesion in English*. London: Logman.
Koselleck R. (1990). *Le future passé. Contribution à la sémantique des temps historiques*, EHESS.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine/Champion.
Mayaffre D. (2008). De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie, *Sémantique & Syntaxe*, 9 : 53-72.
Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles In *JADT 2014*, Paris, 15-32.
Pincemin B. (2006). Concordances et concordanciers. De l'art du bon KWAC, *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, , Albi ed , 33-42. halshs-00356008.
Salem A. « Les séries textuelles chronologiques », *Histoire & Mesure*, vol. VI, n°1/2, 1991, p.149-175.
Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
Skinner Q. (1978), *Foundations of modern political Thought*, 2 vol., Cambridge University Press, 1978
Tournier M. (1997). *Des mots en politique. Propos d'étymologie sociale*. Klincksieck.
Vanni L. et Mittmann A. (2016). Cooccurrences spécifiques et représentations graphiques, le nouveau "Thème " d'Hyperbase. In *JADT 2016*-Nice. pp.295-305. (hal-01359413).

What about corruption? A text analytics method for literature review

Simona Mercurio

La Sapienza University of Rome – simona.mercurio@uniroma1.it

Abstract
Several studies have been conducted worldwide on the phenomenon of a global problem of corruption from the point of view of many aspects of daily life. The aim of this paper is to explore how the corruption has been threated and the context where it has been investigated in national and international literature by a keyword string search method of the term 'corruption' across academic electronic database (i.e. Scopus).
The issue is to study the relationship between words using a method of brand score measurement proposed by SBS (Semantic Brand Score) that combines methods of Social Network Analysis and Semantic Analysis using word co-occurrence network. *'Corruption'* is threated as a brand whose importance is analysed by occurrences in the discourse, its embeddedness in text data and heterogeneity in its text associations; this particular scoping literature review allows to discover a set of socio-economic indicators of corruption and walks the steps towards the operationalization of such a multidimensional phenomenon.

Keywords: corruption; topic modeling; SBS; scoping literature review; network analysis

1. Introduction

One of the most popular and accepted definitions of the phenomenon of corruption is '*the misuse of public office for private gain in a manner that contravenes the rules of the game*'; starting from that definition, plenty of consequences derives from corruption inside the social, political, economic, legal, educational tissue. Among these, lot of studies have associated corruption to inequality because it has been found responsible for losses in GDP growth, (Leite and Weideman ,1999), (Abed and Davoodi, 2000), in the ratio of public education and public health spending to GDP and in many other effects related to GDP value.
In effect, as a consequences of corruption there is the *income inequality* through biased tax systems favoring the wealth and well connected (Gupta et al., 2002), or indirectly, by means of other variables which, in turn, affect inequality, like for example, the level and effectiveness of social spending, (Suryadarma, 2012)

and unequal access to education and public services, with particular regard to health services (Azfar and Gurgur, 2005). OECD has promoted the 'Recommendation on Public Integrity' spotlighting the role of *education* in the fight against corruption. The OECD defines *public integrity* as the 'consistent alignment of, and adherence to, shared ethical values, principles and norms for upholding and giving priority to the public interest over private interests in the public sector'. The role of education in identifying public integrity norms and values and socialising them across society is therefore crucial for facing the collective action problems that enable corruption (Munro and Kirya, 2020).
On the other side, *transparency* is also promoted as one of the most important medicines against corruption. In the recent decade, there has been a massive wave of research and debates about the causes of corruption-driven partly by the growing awareness that corruption is not just a moral problem but also a major impediment to development and growth in large parts of the world (see, for example, Tabellini and Trebbi, 2003).
In this article, we try to demonstrate how the corruption and transparency are linked in scientific literature by a particular scoping review, providing a synthesis of articles published in the last 8 years dealing with corruption- related issues on one side and another synthesis of articles dealing with transparency-related issues on the other side. Given the huge amount of research conducted on the topic, we adopt an innovative methodology, which relies on social network analysis and text mining. We use a composite indicator known as the SBS, Semantic Brand Score, (Fronzetti Colladon, 2018) to assess a brand image and importance through the analysis of textual data (Iezzi et al, 2020), considering dynamic longitudinal trends using data downloaded from a result of a keyword string-search method based on the words '*corruption'* and '*trasparency'* on the electronic wide literature database Scopus. The string-search is applied to the title, abstracts and keywords of the articles and results are downloaded and compacted as a large text file to be analyzed by a web app that implements the SBS. SBS has been conceived to express the importance of a brand along the three dimensions of prevalence, diversity and connectivity, which come from a combination and an innovative use of well-known social network and semantic metrics (Evert, 2005; Wasserman and Faust, 1994).

2. Data collection and analysis

The selection on Scopus is applied to all the articles in English, with no limits in time using the keyword "*corruption*" and, in a next step, the keyword "*transparency*".
The selection applied with no filter and obtained querying 'corruption' on title, abstract and keywords form fields returned about 36,115 documents, instead the same search list on word 'transparency' returns about 130262 documents. Due

to different meanings of the term corruption or transparency the string-search is refined on limited subject areas. Another limitation (choice of articles written in English), could be limiting, nevertheless offers an interesting proxy on world productions written on the brands analyzed. The return list on the word *corruption* is now composed of 20,474 documents and the word *transparency* is cited in 26,876 documents. We assessed the importance of each brand along three clusters, 'education', inequality' and transparency', **at first** in the case of the corpus built on the key search-string '*corruption'* and the clusters 'education', inequality' and 'corruption' **and , secondly**, in the case of the corpus built on the key search-string '*transparency'* , in order to identify how these factors are linked together in the literature and the common topics they are cited in and the association rates between them.

We applied the SBS Brand Intelligence Web App on the corpora created as described above. Files are saved in .csv format and uploaded in the SBS platform; text data is pre-processed to remove stop-words (e.g., 'and'), punctuation and special characters. After pre-processing data, text documents are transformed into a semantic network where nodes are words that appear in the text. An arc exists between a pair of nodes if their corresponding words co-occurred at least once. The SBS App has a parameter file that can be personalized (for example, the window length parameter for the determination of co-occurrences) by the analyst to improve the text processing and make it appropriate in the specific context that is being analyzed. Once the network of co-occurence words is created, the App shows the SBS scores on images, graphs and data reports along the three dimensions of prevalence (the presence of a specific brand in the text), diversity (heterogeneity of the words surrounding a brand) and connectivity (network position of a brand).

2.1. Corpus built around the term 'corruption'

The three clusters are '*education'*, '*transparency'*, '*inequality'*. The SBS time trends interactive line graph shows the dynamic evolution of the semantic brand score for each brand over time.

Figure 1. SBS time trends on corpus built on the search key term of corruption

Figure 1 shows a consistent role of transparency in discussions about corruption. The trend keeps the higher value in 2018. Transparency and the right to access government information are internationally regarded as essential to democratic participation, trust in government, prevention of corruption, informed decision-making, the accuracy of government information, and provision of information to the public, companies, and journalists, among other essential functions in society (Bertot et al, 2010). The Internet has greatly reduced the cost of collecting, distributing, and accessing government information (Roberts, 2006). The power of Internet has given a chance to use of e-government for greater access to information and for promotion of transparency, accountability, and anti-corruption goals. Education is cited with the same intensity as the brand transparency. Table 1 underlies the most associated brands with the word corruption across the time and this could confirm a good orientation about the research of the most pertinent indicators of corruption.

Table 1. Most frequent words and brand associations

Brand	*AssociatedWord*	*Weight*	*Time*	*Unique*
education	**training**	**12.0**	**01/01/2021**	**1**
inequality	**aid**	**12.0**	**01/01/2021**	**1**
transparency	**accountability**	**46.0**	**01/01/2021**	**1**
education	higher	60.0	01/01/2020	0
inequality	income	58.0	01/01/2020	0
transparency	corruption	87.0	01/01/2020	0
education	higher	42.0	01/01/2019	0
inequality	corruption	49.0	01/01/2019	0
transparency	corruption	94.0	01/01/2019	0
education	higher	33.0	01/01/2018	0
inequality	corruption	47.0	01/01/2018	0
transparency	corruption	82.0	01/01/2018	0
education	corruption	46.0	01/01/2017	0
inequality	corruption	31.0	01/01/2017	0
transparency	corruption	66.0	01/01/2017	0
education	**higher**	**60.0**	**01/01/2016**	**1**
inequality	**income**	**30.0**	**01/01/2016**	**1**

Education is coupled with 'higher' and 'corruption, "*Education is one of the most important means to improve personal endowments, build capabilities, overcome constraints and in the process, enlarge available set of opportunities and choices*

for a sustained improvement in well-being". (Venkatraja, Indira, 2011). The higher is the level of education, better is social consciousness.

TOPIC	Topic Words	Weight
T3 - Scientific research on transparency data	study	13189.5
	paper	8871.3
	data	8839.7
T5 – Enforcement government laws	government	10218.9
	public	8637.8
	law	4884.5
T6 – Economic growth	effect	5908.6
	level	4904.5
	findings	4518.7

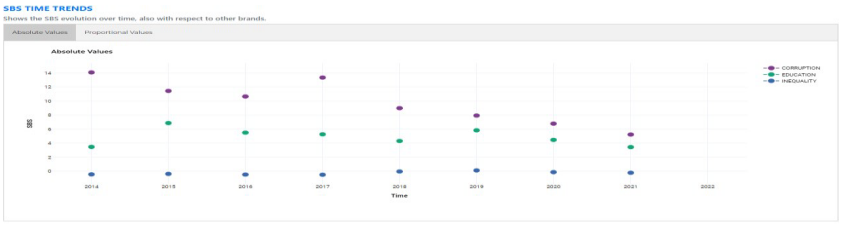
Table 2. Topic Modeling of the first corpus

Interesting results derive from the Topic Modeling that suggests us a strong link between the three clusters in Topic 6 and Topic 5; both topics are placed in laws enforcement and good government strategies (T5) to limit corruption and have a good incidence on economic fields and income equality (T6) and these topics are strongly overlapped in the debates.

2.2. Corpus built around the term 'transparency'

As in the previous paragraph, the core module of SBS has processed a text file whose abstracts talk about transparency following a **parameter** file configured on the three clusters 'education', 'inequality, 'corruption'.

Figure 2. SBS time trends on corpus built on the search key term of transparency



The most cited brand inside literature about transparency are corruption, especially from 2014 and 2018. National and international debates around these

themes have surprisingly increased during last decade. Corruption is strongly associated with transparency and occurs with terms 'government, countries, reduce, anti' et al.

Table 3. Most frequent words and brand associations

Brand	*AssociatedWord*	*Weight*	*Time*	*Unique*
corruption	transparency	53.0	01/01/2021	0
corruption	**anti**	**40.0**	**01/01/2021**	**1**
corruption	countries	33.0	01/01/2021	0
education	**higher**	**42.0**	**01/01/2021**	**1**
education	system	21.0	01/01/2021	0
education	research	20.0	01/01/2021	0
inequality	important	4.0	01/01/2021	0
inequality	economic	3.0	01/01/2021	0
inequality	**algorithms**	**3.0**	**01/01/2021**	**1**
inequality	**distribution**	**3.0**	**01/01/2021**	**1**
corruption	**anti**	**65.0**	**01/01/2020**	**1**
corruption	transparency	53.0	01/01/2020	0
corruption	government	47.0	01/01/2020	0
education	higher	72.0	01/01/2020	0
education	research	26.0	01/01/2020	0
education	quality	24.0	01/01/2020	0

To improve the brand transparency, SBS enforces the idea on the government policies based on the openness and clarity of the public action.

Table 4. Topic Modeling of the second corpus

TOPIC	Topic Words	Weight
T1 – political field	government public	5548.8 4580.8
T2 – policies about transparency	transparency information	20341.4 8160.0
T7 – Economical Field	disclosure firms	2154.7 2127.7

Topics about governance policies about how to reduce and prevent corruption are linked from political and economical point of view (T1 and T7). Technological know-how and expertise (education) can be used to deal with and learn about the effectiveness of transparency strategies (such as anti-corruption). The contact points between Topic are T1 and T2 (governance).

3. Discussion and conclusions

SBS BI App represents an innovative tool to measure brand importance and brand positioning but it is not restricted to brands. This measure can be applied to any set of words to evaluate their importance and recurrent associations. The analysis proposed in this paper is only a start point of different explorations inside the digital academic literature. The use as 'brands' of social factors often linked to corruption, (i.e. lobbying, civism, clientelism) offers the opportunity to discover how studies and debates have discussed across time around these topics, investigating their associations and cause-effects relationships and enforcing the research towards indicators of the phenomenon of corruption. Such indicators can be considered to calculate a new measure of corruption following mathematical procedures that synthetizes multiple information in only one: a multidimensional phenomenon as a unique synthetic measure.

References

Fronzetti Colladon A. (2018). The semantic brand score. *Journal of Business Research*, *88*, 150-160.

Fronzetti Colladon A. and Grippa, F. (2020). *Brand Intelligence Analytics. Digital Transformation of Collaboration*. A. Przegalinska, F. Grippa and P.A. Gloor, eds. Springer Nature Switzerland. 125–141. DOI:https://doi.org/10.1007/978-3-030-48993-9_10.

Evert S. (2005). The statistics of word cooccurrences: word pairs and collocations.

Wasserman S. and Faust K. (1994). Social network analysis: Methods and applications.

Iezzi D. F., Mayaffre D. and Misuraca M. (2020). *Text Analytics*. Springer.

Abed G. and Davoodi H. (2000). Corruption, Structural Reforms and Economic Performance in the Transition Economies. *IMF Working Paper* No. 132. Washington: International Monetary Fund.

Leite C. and Weidman J. (1999). *Does mother nature corrupt? Natural Resources, Corruption, and Economic Growth*, Working Paper WP99/85. Washington DC: International Monetary Fund.

Venkatraja B. and Indira M. (2011). Role of education in social development: an empirical analysis. *Madhya Pradesh Journal of Social Sciences*, *16*(1), 1-11.

Suryadarma D. (2012). How corruption diminishes the effectiveness of public spending on education in Indonesia. *Bulletin of Indonesian Economic Studies*, *48*(1), 85-100.

Roberts A. (2006), *Blacked out: Government secrecy in the information age*, Cambridge University Press, New York

Recanatini F., Prati A. and Tabellini G. (2005). Why are some public agencies less corrupt than others? Lessons for institutional reform from survey data. In *Sixth IMF Jacques Polak Annual Research Conference on Reforms, International Monetary Fund, Washington, DC, November* (pp. 3-4).
Gupta S., Davoodi H. and Alonso-Terme R. (2002). Does corruption affect income inequality and poverty?. *Economics of governance*, *3*(1), 23-45.
Azfar O. and Gurgur T. (2008). Does corruption affect health outcomes in the Philippines?. *Economics of Governance*, *9*(3), 197-244.
Bertot J. C., Jaeger P. T. and Grimes J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government information quarterly*, *27*(3), 264-271.
Munro C. and Kirya M. (2020). Values education for public integrity. U4 Issue 2020:8, *Anti-Corruption Resource Center*, CMI Chr Michelsen Centre.

Using Emotional Text Mining to explore the Spanish culture of Organ Donation

Silvia Monaco, Anna Cortellino, Francesca Greco, Michela Di Trani

University of La Sapienza – silvia.monaco@uniroma1.it
cortellino.1798009@studenti.uniroma1.it michela.ditrani@uniroma1.it
francesca.greco@uniroma1.it

Abstract
Organ donation is the principal way to act the life-saving treatment of transplantation and Spain represents the best practice in this field. The "Spanish model" in organ donation is spread in different countries, but with lower results. Probably other elements can affect the primacy of Spain in this field. More of all, culture can be taken into consideration as the assumption of values on a specific theme, showing how it is emotionally categorized. In this paper, we analyzed two of the most popular Spanish newspapers through Emotional Text Mining (ETM), an unsupervised and bottom-up procedure. We collected all the articles (n= 517) published between 2001-2021 containing the keyword "donacion de organos". Results show a factorial space composed of 5 clusters and 4 factors. The first factor, "The Process", designates a circle, from the family to the patient. The second one, the "Spanish Healthcare", describes the excellence and the State. The third one describes the "Transformation", between the public progress and keep-saving life. The last factor talks about the "Demand-Offer Model". The first cluster, "Person as ONT's principle", underlines the important role of the single citizen. The second cluster represents the "Citizen as center of Institution". The third cluster talks about the "Spanish model in the world". The fourth cluster highlights the "Donation as national project". Finally, the fifth cluster represents "The time problem". The application of ETM in this work showed important representations of the culture of organ donation in Spain.

Keywords: organ donation; emotional text mining; newspapers; culture; Spain

1. Introduction

Transplants are the most effective therapy "for most end-stage organ failure" (López et al, 2018) and organ donation is the first fundamental step in this process, taking into consideration that from one donation eight lives can be saved (OrganDonor.org, 2022). Despite the increase of donations and transplants in the years, a huge gap between the number of available organs and people in need is still present (Lewis et al., 2021) and this is the leading cause of death among

patients waiting on waiting lists (Transplant Observatory, 2022). Organ procurement is a complex process where the most significant final point is the will of the family to donate, or not, the organs of a deceased relative (Matesanz & Miranda, 1998). Spain, reaching 40.2 per million people (p.m.p.) donors in 2021 (IRODAT, 2022), is considered the reference country for organ donation in the world. This primacy is held for many years and has a complex story behind it that bought to the building of the "Spanish Model" (Matesanz & Miranda, 2002). This model starts with the creation, in 1989, of the official agency of Organización Nacional Transplantes (ONT) to face the issue of organ shortage. The fundamental vision was that the problem was not the absence of donors, but the failure to convert potential citizens to donors. The ONT created a complex transplant coordination network, composed of experts, that works locally, regionally, and nationally. The Spanish Model is founded on improving the quality of public information available about organ donation and on the network of this complex organization (Rosenblum et al., 2012). Moreover, the training given to health professionals on "maximizing donation by identifying the donor", the figure of a coordinator working together with his team in each hospital, the direct contact with the ONT, which guarantees and helps the identification of donors, are important facets of the Spanish Model. Since its institution, in Spain there has been an increase of 14.3 p.m.p. donors. Other countries, such as Italy, applied this model in their system, but with lower success. Italy reached only 22.9 p.m.p. donors in 2021, with 28.6% opposition from families to donation (Centro Nazionale Trapianti, 2022). If compared to Spanish one, these data, are very distant from the Spanish results, even if produced by the application of the same model: only 17% of Spanish families opposite to donate in 2021 (ONT, 2022). To understand the deep factors which may influence these differences, citizen's social representation of organ donation must be taken into consideration (Scandroglio et al., 2011). These representations describe how a specific social group (e.g. Spanish citizens) interprets, shares and consequently acts on certain issues (Moscovici, 2005). This study aims to explore the Spanish representations on organ donation, in order to catch the deep social elements which support the primacy of Spain in this field.

2. Method

A sample of 517 newspaper articles from the main Spanish newspapers, 'El País' and 'El Mundo', was used in this research. Each article was selected considering a specific temporal range (2001-2021) and using the keywords "donacion de organos" (trad.= organ donation) during the research in the online database of newspapers. To the final large corpus (Tokens= 316067) was applied the methodology of Emotional Text Mining (ETM; Greco, 2016; Greco & Polli, 2020) with the aid of the software T-Lab (Lancia, 2017). This methodology is

composed by a structured and reproducible procedure and has already been used by several authors in the field of psychology (Cordella et al., 2014; Greco, 2016; Monaco et al., 2021). ETM is a methodology based on a semiotic approach which follow a non-supervised bottom-up procedure for the text mining. It allows the study of the emotional dimensions that organize the relationship between context and individual. ETM preprocess data (lemmatization and corpus fragmentation in context units) and calculate three lexical indexes to assess the richness of the corpus and the possibility to statistically process data (token, hapax percentage, type/token ratio) (Giuliano and La Rocca, 2010). We removed the stopwords, the terms"organ" and "donation" and those with a frequency lower than 10 occurrences. Then, we performed a cluster analysis with a bisecting *k*-means algorithm based on cosine similarity on the term-context units matrix to identify the representations of organ donation. Finally, we calculated the intraclass correlation coefficient to choose the optimal solution, and we performed a correspondence analysis on the term-cluster matrix to identify the symbolic categories setting the communication on organ donation (Greco, 2016).

3. Results

The large size corpus made of 517 newspaper articles showed a good lexical richness (token= 316,067; type/token ratio = 0.067; Hapax = 38%). The corpus was segmented in 6692 context units and the selection of 1,449 terms allowed the classification of 99.5% of the corpus. The optimal partition identified by means of the intraclass correlation coefficient was 5 clusters which are located in a factorial space made by 4 factors (Fig.1).

The first factor (31.5% of inertia), "The Process", designates a circle starting from an active family in the organ donation and ending in the patient waiting for a transplant. This factor underlines the mutual mechanism of Spanish donation: assuming that there are still patients on list, then this produces a participant family that talks about the issue and takes decisions together. The second factor (28.6% of inertia), "Spanish healthcare", describes a State near its citizens and, at the same time, an example of excellence in the healthcare systems. The third factor (23.4% of inertia), "Transformation", shows the renewal of Spanish society between public progress and the maintenance of the circle of life through the process of donation and transplantation. This factor underlines on one hand the will of Spain to continue an improvement of itself, even if it is actually considered the world's excellence in organ donation, and on the other the transformation of people's life thanks to donation. The fourth factor (16.5% of inertia), "Demand-Offer Model", shows a specific lens of lecture of the issue, focused on the resolution of the problem of the waiting list (demand) through donors (offer), which is still present in Spain. This is a more detached vision of organ donation, where data lost the specificity of individuals, allowing a meta-

vision of big data. In this factorial space, 5 clusters are placed with specific coordinates (Tab.1 and Fig. 1).

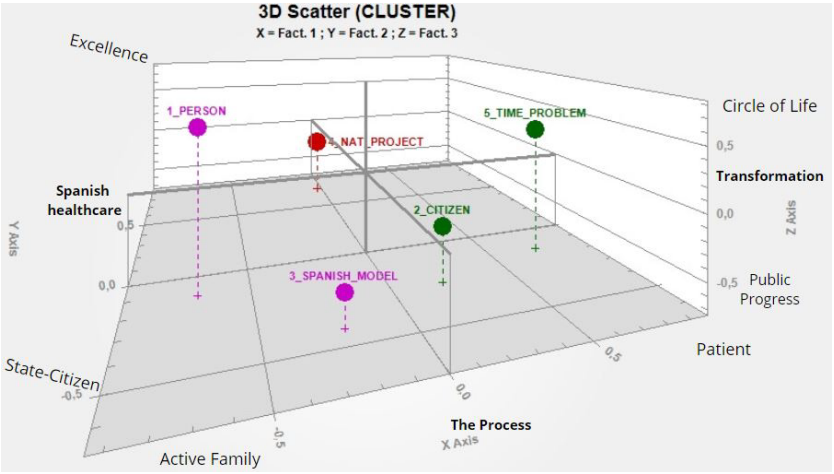


Figure 1. Cluster location in the symbolic space
(the percentage of explained inertia is reported under each factor; the axes coordinates are indicated between brackets for each cluster)

Table 1. Cluster location in the symbolic space
(the percentage of explained inertia is reported under each factor; the axes coordinates are indicated between brackets for each cluster)

Cl	Cluster Label	Factor 1 31,5% The Process	Factor 2 28,6% Spanish Healthcare	Factor 3 23,4% Transformation	Factor 4 16,5% Demand-Offer Model
1	Person as ONT's principle	Active Family (-0.7)	State near its citizens (-0.1)	Circle of Life (0.6)	Offer (0.1)
2	Citizen as center of Institution	Patient (0.2)	State near its citizens (-0.3)	Public Progress (-0.3)	Offer (0.7)
3	Spanish model in the world	Active Family (-0.2)	State near its citizens (-0.4)	Public Progress (-0.5)	Demand (-0.4)
4	Donation as a national project	Active Family (-0.1)	Excellence (0.7)	Public Progress (-0.2)	Demand (-0.1)
5	The time problem	Patient (0.7)	State near its citizens (-0.2)	Circle of Life (0.4)	Demand (-0.2)

The first cluster (18.89% of the total C.U.), "Person as ONT's principle", highlights the important role of every single citizen in the Spanish organ donation system. Individuals are as the recipient as the donor, so they are the starting and ending point of this process which brings to life. The second cluster (13.57% C.U.), "Citizen as center of Institution", highlights the close relationship between the Spanish State and its citizens. In particular, the real final aim of each statal institution is the citizens, and this is interiorized from them. The third cluster (20.74% C.U.), "Spanish model in the world", explains the process of dissemination of Spanish best practices in the world. This cluster underlines an open vision of the donation issue and the Spanish model is not relegated to Spain itself, but it is shared with the world. The fourth cluster (25.69% C.U.), "Donation as a national project", highlights that this issue is considered a national aim, involving a huge amount of resources through the network related to organ donation. Finally, the fifth cluster (21.11% C.U.), "The time problem", represents the dramatic situation of those who are on the waiting list. This cluster underlines the anguished feeling of waiting, among many others, for that piece that will save you.

4. Conclusion

In conclusion, ETM can be considered a valid method to identify cultural dimensions that influence organ donation choices. The use of text mining in public health could be an effective strategy to promote well-being. In this analysis it is possible to assume that the social representations of organ donation explain the cultural meaning of this issue in the Spanish citizens, having an impact on the final rate of organ donor. In fact, the textual analysis of Spanish newspapers has identified areas related to the sense of community and unity, where the organ donation belongs to everyone. Moreover, the results underline an important relationship between the citizens and the State. The analysis of cultural representation on such complex issues, as organ donation, could help to understand specific facets of each country helping to improve the final donor rates. In particular, the Spanish Model works on a terrain where citizens feel the State as an active part of their life, with a good relation between them. Maybe, the huge sense of community is a product of this connection. For the future, other research with the same methodology in other countries could underline differences or similarities in organ donation, helping in this way the construction of focused campaigns for citizens or applying results to the construction of specific training for health workers in this field.

References

Bolasco S. (2013). *L'analisi automatica dei testi: fare ricerca con il text mining*. Carocci Editore.

Centro Nazionale Trapianti. Accessed in 5/03/2022 through the website: https://www.trapianti.salute.gov.it/imgs/C_17_cntPubblicazioni_463_allegato.pdf

Cordella B., Greco F. and Raso A. (2014). Lavorare con un corpus di piccole dimensioni in psicologia clinica: Una proposta per la preparazione e l'analisi dei dati. In Née E., Daube M., Valette M. and Fleury S., eds., *Actes JADT 2014, 12es Journées internationales d'Analyse Statistque des Données textuelles, Lexicometrica*, pp. 173-184
Giuliano L. and La Rocca G. (2010). *Analisi automatica e semi-automatica dei dati testuali: Strategie di ricerca e applicazioni*, Vol. II. Led.
Greco F. (2016). *Integrare la disabilità: Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.
Greco F. and Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management*, vol. 51: 101934.
IRODAT. International Registry In Organ Donation And Transplantation, accessed in 5/03/2022 through the website: https://www.irodat.org/?p=database&c=ES#data
Lancia F. (2017). *User's Manual: Tools for text analysis, T-Lab version Plus*. Available in: https://www.tlab.it/features/
Lewis A., Koukoura A., Tsianos G.I., Gargavanis A.A., Nielsen A.A. and Vassiliadis E. (2021). Organ donation in the US and Europe: The supply vs demand imbalance. *Transplantation Reviews*, vol. 35(2): 100585.
López J.S., Martínez J.M., Soria-Oliver M., Aramayona B., García-Sánchez R., Martín M.J. and Almendros C. (2018). Bereaved relatives' decision about deceased organ donation: An integrated psycho-social study conducted in Spain. *Social Science and Medicine*, vol. 205: 37-47.
Matesanz R. and Miranda B. (1998). International figures on organ donation & transplantation activities 1997. *Newsletter Transplant*, vol. 3 (1), 5–11.
Matesanz R. and Miranda B. (2002). A decade of continuous improvement in cadaveric organ donation: the Spanish model. *Journal of Nephrology*, 15: 22–28
Monaco S., Greco F., Cordella B. and Di Trani M. (2021). Organ donation: a study of its representations among organ procurement coordinators and their staff. *Psicologia della Salute*, vol. 3: 125-142.
Moscovici S. (2005). *Le rappresentazioni sociali*. Il Mulino.
Organ Donor. Accessed in 5/03/2022 through the website: https://www.organdonor.gov/learn/organ-donation-statistics
Organización Nacional Transplantes. Accessed in 5/03/2022 through the website: http://www.ont.es/Documents/Balance%20de%20actividad%20Donaci%C3%B3n%20y%20Trasplante%202021_ONT.pdf
Rosenblum A.M., Horvat L.D., Siminoff L.A., Prakash V., Beitel J. and Garg A.X. (2012). The authority of next-of-kin in explicit and presumed consent systems for deceased organ donation: an analysis of 54 nations. *Nephrology Dialysis Transplantation*, vol. 27(6): 2533-2546.
Scandroglio B., Domínguez-Gil B., López J.S., Valentín M.O., Martín M.J., Coll E., Martínez J.M., Miranda B., San José M.C. and Matesanz R. (2011), Analysis of the attitudes and motivations of the Spanish population towards organ donation after death. *Transplant International*, vol. 24: 158-166.
Transplant Observatory. Accessed in 5/03/2022 through the website: http://www.transplant-observatory.org/

Questionner la posture énonciative d'un journal à visée humoristique: étude textométrique et contrastive

Cyrielle Montrichard

Université de Neuchâtel – cyrielle.montrichard@unine.ch

Abstract
This contribution explores with textometric tools the enunciative stance of an online French newspaper which copies the style of traditional newspapers with a humorous purpose: *Le Gorafi.* The hypothesis of this work is that despite its reproduction in *the way of saying*, we can observe an enunciative stance that is very different from what we find in traditional newspapers. To point out these differences, we use statistical tools to question a corpus containing both articles from *Le Gorafi* and from *Le Figaro* (French newspaper).

Keywords: textometry, syntactic pattern, enunciative stance, détournement

Résumé
Cette contribution vise à interroger, à l'aide des outils textométriques, la posture énonciative d'un journal à visée principalement humoristique : *Le Gorafi.* Par une mise en contraste avec le journal dont *Le Gorafi* revendique le détournement (*Le Figaro*), nous proposons de mobiliser deux outils statistiques pour voir émerger les différences de postures énonciatives : l'indice de spécificité et une AFC originale fondée sur des patrons syntaxiques.

Mots clés : textométrie, patrons syntaxiques, posture énonciative, détournement

1. Introduction

Ce travail interroge, à l'aide d'une démarche outillée et du logiciel TXM (Heiden et al.,2010), les caractéristiques énonciatives d'un titre de presse en ligne à la trajectoire particulière et aux visées singulières : *Le Gorafi*. Ce journal affiche une ligne éditoriale qui vise au détournement du genre en produisant des articles qui s'inscrivent dans le genre journalistique mais qui se fondent sur ce que nous proposons d'appeler des *fausses nouvelles* :

> Tous les articles relatés ici sont faux (jusqu'à preuve du contraire) et rédigés dans un but humoristique. L'utilisation de noms de personnalités ou d'entreprises est ici à but

purement satirique.
(In « À propos » sur le site du Gorafi : https://www.legorafi.fr/about/)

Cette ligne éditoriale ne s'inscrit dans aucune des catégorisations traditionnelles. Si *Le Gorafi* rejoint dans une certaine mesure la presse satirique, il s'en distingue largement par l'absence totale d'une visée d'information – qui, selon Patrick Charaudeau (2011), associée à la visée de captation, constituent la finalité du contrat de communication médiatique. De plus, cette presse en ligne – qui en reprend les caractéristiques formelles (rubriquage, titre, chapeau, *etc*.) – est également affranchie des contraintes liées aux « conditions de vérités » (puisque le journal revendique la fausseté des informations communiquées), conditions intrinsèques au contrat médiatique originel (Charaudeau, 2011) et pesant sur la presse traditionnelle.
L'objectif de cette étude est d'interroger avec les outils textométriques la posture énonciative d'un journal à visée principale humoristique lorsque ce dernier s'inscrit dans *une manière de dire* qui se veut conforme à celle de la presse écrite traditionnelle mais qui s'inscrit en marge de cette dernière en termes de visées et de contraintes. En effet, nous postulons que les éléments extratextuels (le contrat de communication) pesant sur la situation d'énonciation modifient la posture énonciative du journal humoristique bien qu'il vise à imiter et s'inscrire dans le genre journalistique. La mise en contraste du *Gorafi* avec le journal qu'il revendique imiter (*Le Figaro*), permettra de faire émerger des éléments linguistiques liés à la posture énonciative des titres et de les distinguer.
Nous proposons d'abord de revenir plus finement sur le contexte de la recherche qui présente le cadre théorique d'analyse et les données textuelles étudiées. Nous présenterons ensuite les méthodes, résultats et analyses lors de l'exploration textométrique des deux journaux.

2. Contexte de la recherche et corpus

Saisir la posture énonciative d'un journal est complexe et intrinsèquement liée à l'identité du titre. À l'instar de Roselyne Ringoot (2015 : 18), nous considérons que cette dernière peut être appréhendée par l'étude d'une part de « l'énonciation éditoriale » (identité visuelle, rubriquage) et d'autre part de « l'énonciation textuelle » (le contenu rédactionnel). La première permet de tisser un lien étroit entre *Le Gorafi* et *Le Figaro.* En effet, plusieurs indices tels que le jeu sur le nom du titre, l'identité visuelle similaire (le logo, les couleurs), la *filiation* entre les deux titres mentionnée dans la section « à propos »[1] confirment que *Le Gorafi*

[1] On peut notamment y lire : « Le Gorafi est né après un conflit d'intérêts avec les créateurs du Figaro en 1826. »

s'inscrit dans un rapport de détournement du *Figaro*. L'étude du contenu rédactionnel, à la recherche des traces de l'énonciateur, permettra de mettre en lumière des différences dans la posture énonciative des titres, malgré un discours de part et d'autre qui reprend les codes et manières de dire du genre journalistique.

Genre à visée principalement informative et affichant une neutralité objective, le discours journalistique est un des lieux privilégiés pour étudier les jeux d'effacement énonciatif (EE) (Koren, 2004). L'énonciateur (que nous associons au journaliste/à la voix mettant en scène les autres voix) orchestre le point de vue de locuteurs et affiche une neutralité objective notamment dans le sous-genre de l'article d'information – nous excluons ici les sous-genres journalistiques qui assument (plus ou moins) une prise de position tels que l'éditorial ou la chronique.

Derrière ce jeu énonciatif de suppression apparente de la voix derrière l'information communiquée, des études ont notamment cherché à débusquer les traces de la présence de l'énonciateur (Koren, 2004). Les travaux d'Alain Rabatel (2004a et 2004b), quant à eux, ont questionné la posture de l'énonciateur, apportant des clés essentielles pour la caractériser. En effet, il propose de distinguer trois types de posture face à l'interlocuteur : l'énonciateur peut s'inscrire dans la co-énonciation (dans la recherche d'un consensus), la sous-énonciation (le point de vue est dominé) ou la surénonciation (le point de vue se veut dominant et surplombant) et ce positionnement énonciatif se donne à lire dans l'énoncé même (Rabatel, 2004a : 9). Cette contribution propose d'interroger spécifiquement la posture énonciative du *Gorafi,* par-delà l'EE, au prisme d'un cadre théorique à la jonction entre énonciation, analyse du discours et argumentation en mobilisant les outils textométriques. Notre corpus de travail est fondé sur la mise en série d'articles du *Gorafi* parus entre 2012 et 2015 et s'inscrivant dans les rubriques *Société*, *Sciences*, *Sport, Culture, Monde* et *Politique.* Nous avons aussi intégré des articles du *Figaro* publiés entre 2012 et 2013 prenant place dans des rubriques équivalentes[2]. Le corpus compte 531'733 occurrences et des parties équilibrées : 264'646 occurrences pour *Le Figaro* et 267'087 pour *Le Gorafi*. Nous décrirons d'abord les résultats les plus saillants obtenus à l'étude des unités linguistiques participant à la monstration d'une posture énonciative et repérées avec l'indice de spécificité (Lafon, 1980).

[1] Un rubriquage – lieu où se construit aussi l'énonciation éditoriale selon Ringoot (2015) – similaire est observé entre Le Gorafi et Le Figaro. Au sein des rubriques sélectionnées, seule la rubrique Monde du Gorafi se distingue par son intitulé de la rubrique International du Figaro – mais nous les avons considérées comme équivalentes.

Ensuite, nous montrerons comment nous avons interrogé les patrons syntaxiques à l'aide de l'analyse factorielle des correspondances (AFC), considérant, à l'instar de Rabatel (2004a et 2004b), que ces structures peuvent révéler une posture énonciative.

3. Indice de spécificité sur une sélection de catégories morphosyntaxiques

Cette partie explore les lemmes en spécificité positive dans le *Gorafi* (contrasté avec le *Figaro*) et plus spécifiquement ceux issus de trois catégories morphosyntaxiques qui sont pertinentes dans l'étude de la posture énonciative : les adverbes, les pronoms, et les verbes. Au sein de la catégorie des adverbes, les intensifieurs qui sont, pour Anscombre et Ducrot (1983), des indices d'un point de vue sur le dit indiquant la direction argumentative de l'énoncé, ont retenu notre attention. Ce type d'adverbes peut aussi être associé, selon nous, à l'idée de « renchérissement », signe d'une surénonciation pour Rabatel (2004a : 12). On notera ainsi l'usage en sur-emploi de « très » (indice de spécificité de +39,5), « tout » (+13), « vraiment » (+11,5), « trop » (+7,5), « beaucoup » (+7), « absolument » (+2,5), « exclusivement » (+2), « tellement » (+2). Ces adverbes (en nombre dans l'extrait (1) ci-après) sont également, selon nous, à mettre en lien avec la posture satirique et parodique du journal qui caricature le discours journalistique – et donc force le trait. On trouve également des adverbes marquant une modalité en spécificité positive significative : « peut-être » (+13), « évidemment » (+4), « personnellement » (+3), « franchement » (+2,5). La modalité (*modus*), quand celle-ci est incorporée – au moyen d'un adverbe – au *dictum* (Bally, 1965), est un indice d'une position « en surplomb » d'un énonciateur qui par un dédoublement énonciatif du même temps qu'il dit, commente le dit (Vion, 2004 : 102). Le deuxième élément marquant concerne l'emploi des pronoms en spécificité positive[3]. Nous notons particulièrement l'usage significatif des pronoms « on » (+19) et « ce » (+69.8) pouvant être associé à une posture énonciative en surplomb (Rabatel, 2004b) par leur capacité à désinscrire l'énonciateur de son dire et à jouer sur le flou du référent – voir extrait (1). Le dernier élément émergeant de l'étude, sur lequel nous souhaitions insister, est la place prépondérante des verbes. Dans la liste des verbes sur-employés, de nombreux *verbes de dire* sont aussi relevés et peuvent aussi être un indice de surénonciation lorsqu'ils introduisent un discours rapporté (Rabatel, 2004a). Trois usages de ces verbes sont observés en contexte : (i) ils introduisent une parole à visée illustrative (« raconter » (+19), « commenter » (+18.5)) ; (ii) ils qualifient et imposent un cadre d'interprétation du dire rapporté

[3] Nous n'aborderons pas l'usage du « je » (+136) qui est surtout l'indice d'un recours fréquent aux citations.

(« affirmer » (+11.5), « annoncer » (+10), « décider » (+10), « avouer » (+7)), enfin, (iii) ils rapportent au discours indirect libre les pensées et désirs des actants se posant ainsi en narrateur omniscient (« refuser » (+12), « souhaiter » (+7.5)). Ces trois usages dans le corpus peuvent, selon nous, s'inscrire dans une posture énonciative surplombante puisque : dans le premier cas, le discours rapporté n'est qu'un support de l'argumentation, de la proposition ; dans le deuxième cas, en imposant un cadre d'interprétation du discours rapporté, l'énonciateur guide les lecteurs et ainsi impose son point de vue sur le dire ; et dans le dernier cas, l'énonciateur se présente comme un narrateur omniscient rapportant au discours direct libre la parole mais aussi les désirs des locuteurs. L'extrait ci-dessous illustre les cas (i) et (ii).

> (1) […] « My Little Pony » a rapidement conquis un public large, […]. Un phénomène qui commence à **inquiéter** dans les **plus** hautes sphères de l'Etat. […] « Si **on** ne fait rien, il sera **trop** tard, […] » **affirment** les psychologues dans un premier rapport préliminaire. […]
> En outre, la **trop** forte identification de certains fans entraîne une perte de repères au quotidien qui peut **très** vite devenir un obstacle dans la vie de tous les jours. **«** Il y a des cas de gens qui sont partis **trop** loin dans leur passion, il suffit de voir les images de ces conventions, **c**'est **absolument** effrayant » **commente** un parent. « J'ai interdit à mon enfant de regarder cette série, je ne veux pas qu'ils deviennent comme **ces** gens » **témoigne** un autre parent, inquiet. (*Gorafi*, 2012)

Cette étude des lemmes spécifiques étiquetés comme verbe, adverbe ou pronom a permis de montrer un faisceau d'éléments qui semble indiquer un posture énonciative surplombante : adverbes de modalité, intensifieurs, verbes de dire modalisant le discours rapporté et usage de pronoms indéfini et démonstratif. Il s'agit désormais de voir si cette posture surplombante peut aussi s'observer dans l'étude des patrons syntaxiques.

4. Exploration des structures syntaxiques au moyen de l'AFC

L'AFC ci-dessous est fondée sur la distribution des 50 patrons syntaxiques de longueur trois unités les plus fréquents[4] dans le corpus partitionné par rubrique et journal. Cette méthode permet d'observer précisément la consistance du discours dans un rapport syntagmatique. Nous avons mis en évidence sur l'AFC les patrons syntaxiques contribuant le plus fortement à la structure de l'axe 1 – voir aussi le tableau 1 ci-après. Sur cet axe (taux d'inertie de 43,34%), l'opposition par journal s'observe avec à droite de l'axe, du côté du *Gorafi,* des

[4] Nous avons exclu des patrons interrogés les marques de ponctuation (faible, forte et de citation).

patrons syntaxiques à consistance verbale et, à l'inverse, à consistance nominale du côté du *Figaro*. Si une des caractéristiques de l'EE est la prégnance d'un discours à consistance nominale (Rabatel, 2004a), force est de constater que celui-ci se situe du côté du *Figaro,* indice de l'absence marquée de l'énonciateur au profit d'un point de vue exprimé implicitement. L'analyse voit émerger des points importants quant aux structures qui permettent de mieux caractériser la posture énonciative de chacun des titres.

Les rubriques du Figaro, à gauche de l'axe (exceptée la rubrique sport), voient à leur côté un patron syntaxique représentatif d'un discours à consistance nominale (n°2 dans le tableau 1). Une des particularités du Figaro est d'user de syntagme nominal pouvant être le signe d'un effacement énonciatif par une « objectivation de l'objet » (Rabatel, 2004a : 4) comme dans « La paralysie de la communauté internationale devient de plus en plus intenable » (Figaro, 2012) où la paralysie est présupposée.

Le Gorafi, dont toutes les rubriques se situent à droite de l'axe, est du même côté que les patrons syntaxiques n°1 et n°3 qui renvoient à des constructions souvent impersonnelles. En observant en contexte le segment répété « il y a » (n°3), on relève une construction argumentative fondée sur la propriété linguistique du patron syntaxique qui, lorsque introduisant une relative, permet de renforcer l'assertion de la « relation établie dans le domaine du deuxième verbe » (Blanche-Benveniste, 1982 : 97).

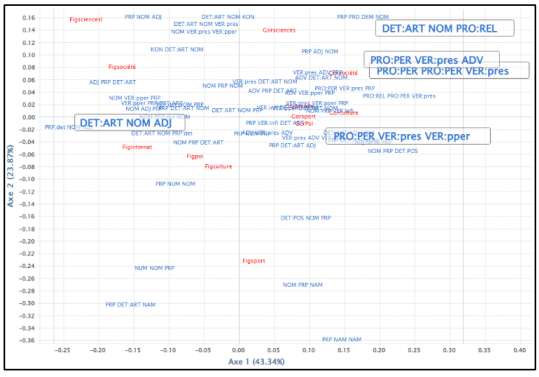


Figure 1. AFC réalisée sur les 50 premières constructions syntaxiques de trois unités (en bleu) sur le corpus partitionné selon les rubriques des journaux (en rouge)

Tableau 1. Structures syntaxiques participant à la structure de l'axe et exemples en corpus

n° (id)	Patrons syntaxiques	côté de l'axe	Exemples dans le corpus (Fréq. Brute)
1	pronom personnel | verbe | adverbe	droit	« y a pas » (19), « y a beaucoup » (8), (sous-corpus *Gorafi)*
2	déterminant article | nom | adjectif	gauche	« la semaine dernière » (23), « l'an dernier » (22), (sous-corpus *Figaro*)
3	pronom | pronom | verbe	droit	« il y a » (142), « il s'agit » (121) (*Gorafi*)
4	déterminant | nom | pronom relatif	droit	« Une annonce qui » (13), « Une situation qui » (10), (*Gorafi*)

Ainsi, dans l'extrait (1), la construction avec « il y a » asserte la relation entre « des cas de gens » et « sont partis trop loin ». Ce type de structure tend donc d'une part, à faire valoir un point de vue difficile à réfuter et d'autre part, à afficher une objectivité – une neutralité descriptive est perceptible avec « il y a ». On repère aussi, à droite de l'axe, le patron n°4 qui, par une construction à l'aide d'une relative, semble fonctionner à la manière des « propositions relatives explicatives » (Rabatel, 2004b : 31). Dans notre corpus, ce type de construction (exemple en contexte dans l'extrait (1) : « un phénomène qui ») permet à l'énonciateur d'introduire la proposition comme « allant de soi » (*ibid*). En effet, dans l'extrait (1) le lecteur ne questionnera pas *naturellement* l'inquiétude liée au phénomène. Ces types de construction peuvent donc être associés à une argumentation faisant figure d'autorité, liée à une posture de surénonciation visant à dominer l'échange sans pour autant que l'énonciateur prenne en charge l'énoncé.

L'AFC, et la mobilisation des taux de contribution à la structure des axes, permet la mise en lumière des patrons syntaxiques singularisant la manière de dire des deux titres. Par ailleurs, elle représente, non plus seulement l'axe paradigmatique mais l'axe syntagmatique offrant ainsi des perspectives intéressantes dans l'objectif de « [con]textualiser » les formes et les discours (Mayaffre, 2014 : par.2).

5. Bilan et perspectives

Ce travail proposait de poser un regard sur la posture énonciative de deux journaux aux lignes éditoriales distinctes : *Le Gorafi* et *Le Figaro.* Cette recherche visait avant tout à percevoir, à travers une étude contrastive centrée sur la posture énonciative, les traces linguistiques qui permettent de distinguer un journal visant à détourner le genre tout en en reprenant les codes.

Mis en contraste, *Le Figaro* et *Le Gorafi* montrent une *manière de dire* très différente tant au niveau du vocabulaire que dans les constructions syntaxiques mobilisées. Montrant un vocabulaire plus neutre, un discours à consistance nominale, *Le Figaro* s'inscrit, sans surprise, dans l'EE – il s'agit d'un premier état d'analyse qui mériterait d'être confirmé dans un corpus intégrant d'autres journaux traditionnels. Les différences relevées semblent surtout liées à la posture énonciative singulière du *Gorafi*. En effet, nous avons montré que ce journal s'inscrit en sur-énonciateur grâce aux repérages de nombreuses traces linguistiques impliquant une telle posture dont : les structures impersonnelles, la dissipation du *modus* dans le *dictum*, l'usage significatif des pronoms indéfini et démonstratif ainsi que l'abondance des verbes de dire et donc de discours rapportés (souvent illustratifs de la proposition ou introduits par des verbes guidant le cadre d'interprétation). Finalement, en fournissant une caricature du *Figaro, Le Gorafi,* affranchi des conditions de vérité et de la restitution des faits

et des points de vue circulant dans l'espace public, construit un sur-énonciateur dont la voix est dominante, unique et campée derrière un EE. Forçant le trait du discours journalistique contemporain, *Le Gorafi* montre un miroir grossissant, et certes aussi déformant, de la pratique d'écriture journalistique.

Bibliographie

Anscombre J.-C. et Ducrot O. (1983). *L'argumentation dans la langue*. P. Mardaga.
Bally C. (1965). *Linguistique générale et linguistique française*. Berne : Francke.
Blanche-Benveniste C. (1982). Examen de la notion de subordination. *Recherches sur le français parlé*, vol.(4): 71-115.
Charaudeau P. (2011). *Les médias et l'information. L'impossible transparence du discours.* De Boeck Supérieur.
Heiden S., Magué J.-P. et Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Bolasco S., Chiari I. and Giuliano L., editors, *Statistical Analysis of Textual Data -Proc.of JADT 2010*.
Koren A. (2004). Argumentation, enjeux et pratique de l'« engagement neutre » : le cas de l'écriture de presse. *Semen,* vol.(17): en ligne. DOI : https://doi.org/10.4000/semen.2308.
Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots,* vol.(1) : 127-165.
Marnette S. (2004). L'effacement énonciatif dans la presse contemporaine. *Langages*, vol.(156) : 51-64.
Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). In Née E, Daube J.-M. *et al., éditeurs, Proc of JADT 2014*, pp.15-32.
Rabatel A. (2004a). L'effacement énonciatif dans les discours rapportés et ses effets pragmatiques. *Langages*, vol.(156), 3-17.
Rabatel A. (2004b). Stratégies d'effacement énonciatif et posture de surénonciation dans le Dictionnaire philosophique de Comte-Sponville. *Langages*, vol.(156): 18-33.
Ringoot, R. (2015). *Analyser le discours de presse*. Paris : Armand Colin.
Vion R. (2004). Modalités, modalisations et discours représentés. *Langages*, vol.(56): 96-110.

Recherche Qualitative : Analyse lexicale et sémantique de grands corpus d'entretiens ou de documents

Jean Moscarola[1], Florence Laval[2], Caroline Mothe[3]

[1]Business Science Institute – jmoscarola@gmail.com
[2]Université de Poitiers – flaval@poitiers.IAE-France.fr
[3]Université Savoie Mont-Blanc – carolinemothe@yahoo.fr

Abstract
The digitization of documents and the use of the Internet offer qualitative research increasingly voluminous corpora. Their analysis presents an opportunity and a challenge highlighted in a recent book entitled Big Quali . Textual data analysis is a response to the approach of these massive corpora. From two examples concerning a corpus of 195 documents on the one hand, and more than 415 semi-directed interviews on the other hand, this article examines the scope and limits of the top-down hierarchical classification methods popularized by the Alceste method. In the presence of large heterogeneous corpora, the division into short units of meaning (segments or sentences) does not allow the semantic unity of the components (each document or interview) to be taken into account to establish a typology. The paper presents how the use of a thesaurus allows to remedy this drawback. After presenting the classical method and its alternatives, they are illustrated on the two cases using the Sphinx software

Key Words: Qualitative research, text mining, semantic analysis, classification, software

Résumé
La numérisation des documents et l'utilisation d'internet offrent à la recherche qualitative des corpus de plus en plus volumineux. Leur analyse présente une opportunité et un challenge mis en évidence dans un ouvrage récent intitulé Big Quali[1]. L'analyse de données textuelles est une réponse à l'approche de ces corpus massif. A partir de deux exemples concernant un corpus de 195 documents d'une part, et plus de 415 entretiens semi dirigés d'autre part cet article examine la portée et les limites des méthode de classification hiérarchique descendante popularisées par la méthode Alceste. En présence de grands corpus hétérogènes le découpage en unité de significations de courte taille (segments ou phrases) ne permet pas prendre en compte de l'unité sémantique des composants (chaque document, ou entretien) pour établir une typologie. Le papier présente comment le recours à un thésaurus permet de remédier à cet inconvénient. Après

[1] Daniel Bo Dunod 2022

présentation de la méthode classique et de ses alternatives, elles sont illustrées sur les deux cas en utilisant les logiciels Sphinx et en exposant toutes les possibilités qu'elles offrent pour la fouille de texte.

Mots clés: Recherche qualitative, fouille de texte, analyse sémantique, classification, software

1. Introduction

Ce papier s'adresse aux chercheurs de la recherche qualitative (Dumez 2016). Ils ont l'habitude du travail de lecture et de réflexion, patient aller et retour entre sources documentaires ou entretiens et théories, au cours duquel la pensée s'organise pour donner naissance à un nouveau texte sur la question étudiée. Cette tradition est mise à l'épreuve par la numérisation des sources documentaires et les opportunités de l'internet qui, en donnant accès à des textes de plus en plus volumineux, peuvent nécessiter le recours aux logiciels d'analyse de données textuelles. Leur usage se heurte cependant à la compréhension des méthodes mises en œuvre et peut être, plus profondément, à la légitimité du détour statistique comme préalable à la prise de connaissance. En effet l'analyse des données textuelles consiste à mettre en évidence les propriétés statistiques (lexicales ou sémantiques) du corpus étudié, comme préalable à toute lecture ou interprétation. L'objectif de ce qui suit et d'expliquer les méthodes de classification automatique, l'usage de thésaurus et la fouille de texte en mettant en évidence les conditions et contraintes de leur emploi, d'une part, et la portée de leur contribution comme aide à la construction du sens, d'autre part. On s'appuie pour cela sur l'exemple de 2 recherches menées en collaboration entre des chercheurs de la tradition qualitative et un spécialiste du logiciel Sphinx.

Liens vers l'ensemble des analyses

https://vip.sphinxonline.net/FLAVALGRH/LaFRH-22816Phrases/Page-Liens.pdf

Ce lien pointe sur l'analyse des cas présentés. Elles illustrent les méthodes et réflexions présentées dans l'article. Les consulter est nécessaire à la bonne compréhension du propos.

2. Deux cas de corpus massif: les réponses de l'analyse de données textuelle.

Le premier porte sur les représentations de la fonction RH recueillies par voie d'entretiens menés par des étudiants de Licence 3. Au total le corpus contient 415 entretiens effectués au cours des 4 dernières années. Dans le second, il s'agit d'identifier la vision des acteurs concernés par les réseaux de chaleurs dans le contexte du changement climatique (Mothe & ali 2020). L'analyse de leur discours est menée à partir de 195 documents numérisés. Ces corpus volumineux de plus de 1000 pages et 500 000 mots chacun, pourraient être lus pour dégager une synthèse ou servir à argumenter les thèses défendues par les chercheurs. Mais la charge de travail que cela représente et la difficulté à garantir

l'objectivité des conclusions résultant d'une telle démarche conduisent à envisager d'autres solutions. L'analyse de contenu systématique (Saladana 2013, Krippendorff 2013) consistant à lire l'intégralité du corpus pour en coder les contenus dans une grille établie à cet effet est une autre voie tout aussi lourde, fastidieuse et discutable du fait de la subjectivité du codage, même contrôlée par double codage. Restent les solutions de l'analyse de données textuelles (Lebart & ali 2021). Elles consistent, dans un premier temps, à remplacer la lecture par une analyse lexicale et statistique consistant à mettre en évidence les structures linguistiques du corpus (regroupement de ses éléments en catégories porteuses de sens) et les influences du contexte (spécificités relatives aux auteurs des documents ou identité des interviewé). Ces méthodes apparues dans les années 70, conduisent à une lecture informatisée réduisant la lecture du corpus, quelle qu'en soit la taille, à celle de substituts produits de manière automatique et instantanée (Moscarola 2018). Ces substituts établissent en quelques pages des propriétés observées de manière objective et reproductible. Les analyses présentées donnent un exemple des résultats obtenus pour chacun des cas. Elles mettent en évidence :

- les thèmes développés dans ces corpus et leur poids. Leur importance visualisée graphiquement est établie par le nombre des éléments du corpus qui s'y rattachent.
- l'influence de l'identité des auteurs ou personnes interviewés sur les thèmes mis en évidence.

2.1. Les entretiens sur la fonction RH.

Le corpus des 415 entretiens a été découpé en phrases(n=22 816). Le lien 1 ci-dessous donne

- un aperçu des contenus avec le nuage des mots les plus fréquents et une typologie des phrases regroupées par similarité et définies par les mots qui les distinguent les unes des autres. Ces éléments mettent en évidence les différentes dimensions de la fonction RH.
- le croisement entre l'identité des répondants et les thèmes dégagés par la typologie. Il montre que les hommes et ceux des grandes entreprise associent la fonction RH au thème « entreprise, salarié et service » alors que les femmes et les gens de 40-49 ans sont plus proche des thèmes « personne, poste travail » et « formation… »

Lien 1 entretiens RH https://cvip.sphinxonline.net/tiny/v/QHA50h4PAU

2.2. Les documents sur les réseau de chaleur et le changement climatique.

Le corpus est analysé globalement en prenant les documents comme unité de signification. La typologie est fondée sur les concepts identifiés par un thésaurus général du français et regroupe les documents selon la similarité des concepts qui

les caractérisent. Le lien 2 met en évidence l'interprétation de la classification selon 3 grands thèmes illustrés par les mots et concepts qui les caractérisent et par le type d'acteur : Le thème dominant « l'injonction » est caractéristique des Agence d'Etat et Associations de promotion, il renvoie aux concepts d'espace, de nature et de technique. Le thème « mise en œuvre » est spécifique des exploitants de réseaux et renvoie à l'action. Celui de la « réflexion » enfin est proche des chercheurs, et privilégie les concepts de temps et activités économiques.

Lien 2 Documents réseau de chaleur https://d.ergole.fr/v/Dp0UPE19F5

Ces exemples illustrent l'apport de l'ADT consistant à exploiter les propriétés du corpus en mettant en évidence les mots qui se répètent, les typologies de phrases ou de documents et les croisement avec l'identité des auteurs. Au chercheur de savoir les interpréter et de revenir au texte pour contrôler ses interprétations.

3. Les méthodes

3.1. Niveau d'analyse

La question du niveau d'analyse est déterminante: à quel niveau observer la formation du sens exprimé dans les corpus: au niveau des auteurs ou des personnes considérées comme exprimant une expression/opinion globale qu'il s'agirait de résumer? Ou au niveau des fragments d'expression en considérant par exemple que chaque phrase et porteuse de sens et que les locuteurs ou auteurs puisent dans des répertoire d'idées indifférenciées qu'il s'agit de découvrir ? Cette discussion renvoie à l'histoire des méthodes d'analyse de données textuelle (Moscarola 2018). Elle a débuté par la statistique lexicale consistant à utiliser l'ordinateur pour identifier et dénombrer les formes graphiques d'un corpus, prises comme indicateurs de sens au niveau le plus fin des mots bruts[2]. L'analyse morpho syntaxique améliore le procédé en soulevant l'ambiguïté des formes brutes par l'identification des lemmes[3] et des catégories grammaticales. Les méthodes de classification complètent cette approche par analyse des cooccurrences qui situent les mots dans le contexte de la phrase révélant ainsi les structures sémantiques du corpus. Enfin l'usage des thésaurus (Dasilva 2006), grâce à la modélisation sémantique, rend possible le passage des mots aux concept pour l'étude d'unités de signification de plus grande taille ou pour mesurer des intensités lexicales (Boughzala & ali 2014) indiquant le poids d'un concept en fonction de celui des mots qui le définissent. Ces évolutions provenant de champs disciplinaires différents (statistique, linguistique, lexicographie, sémantique) ont bénéficié de l'ingénierie logicielle qui les a rendues accessibles en donnant accès aux méthodes classiques de l'ADT, aux outils de l'analyse multivariées, à l'analyse sémantique et à l'infographie dynamique. Les dernières

[2] Dans leurs formes « fléchie » genre, nombre flexion des verbe

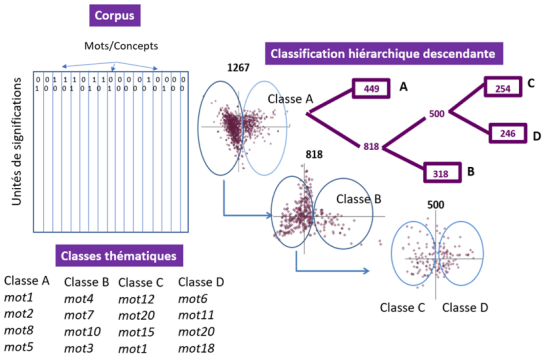
[3] Forme racine d'un mot : singulier, masculin, infinitif

version de Sphinx IQ3 et Dataviv (Moscarola 2018) illustrent bien cette intégration mise en œuvre dans les exemples que nous venons de citer.

3.2. Analyse factorielle multiple et classification hiérarchique descendante

A la suite des travaux de Benzecri, l'analyse factorielle des correspondances multiples a trouvé un champs d'application pour l'analyse des textes traités comme des données qualitatives ou booléenne. D'où l'expression analyse de données textuelles. Reinert en a fait une méthode, connue comme la méthode Alceste, nom du premier logiciel éponyme et reprise ensuite par les logiciels Sphinx Lexica, Iramuteq puis Sphinx IQ et Dataviv (Moscarola 2018). Elle repose sur une séquence d'analyses factorielles multiples (Figure 1) consistant à répartir les unités de significations élémentaires d'un corpus en 2 classes composées d'éléments similaires du point de vue des mots qui les composent. La classe regroupant le plus grand nombre d'unités de signification est répartie à son tour en 2 classes par une nouvelle analyse factorielle. La décomposition se poursuit en divisant à chaque étape la classe la plus nombreuse. Elle s'arrête lorsque la classe de plus fort effectif a un effectif trop faible pour être à nouveau divisée ou lorsqu'un nombre de classes fixé a priori est atteint. Les analyses factorielles multiples portent sur le tableau des occurrences comportant en ligne les unités de significations et en colonnes les mots qui les composent. On obtient ainsi un tableau de 0 et de 1 indiquant la présence au moins une fois des mots décrivant les unités des signification . Cette caractéristique de l'algorithme rend la méthode inappropriée à l'examen de corpus découpé en unité de significations trop longues ce qui reviendrait à considérer comme associés, des mots trop éloignés les uns des autres, ou présents un grand nombre de fois alors que d'autres ne le seraient qu'une fois. C'est la raison pour laquelle cette méthode s'appliquait à l'origine à des unités de significations courtes (séquences de mots arbitrairement fixés 15 à 20, ou phrases).

Figure 1. Classification hiérarchique descendante



3.3. L'analyse sémantique

Pour adapter la classification à des unités de signification longues il convient de construire la table en disposant en colonne non plus les mots mais les concepts qui les caractérisent. Chaque document est alors décrit par le suite des concepts qui lui sont attribués par l'analyse sémantique et l'algorithme de classification débouche sur une typologie des documents ainsi analysés.

Cette manière de procéder a permis de conserver l'idée originelle de classification hiérarchique en la faisant bénéficier des apports de l'analyse sémantique. Elle repose sur une modélisation préalable du sens mise en œuvre par un thesaurus, des ontologies et des réseaux sémantiques. Ils permettent d'opérationnaliser de manière informatique le fonctionnement du triangle sémiotique (Moscarola 2018): Le thesaurus organise «les signifiants» (le sens) en une arborescence d'idées principales, de concepts et de concepts détaillés. A chaque feuille de l'arbre sont rattachées les ontologies, listes de mots, «les signifiants» qui définissent le concept correspondant. Enfin un réseau sémantique, «le référant» précise le contexte des mots selon les idées auxquelles ils sont rattachés.

3.4. L'apport des thésaurus. Thésaurus générique et thésaurus ad hoc

En résumant chaque unité de signification par une liste de concepts, l'usage d'un thésaurus général[4] donne une idée rapide des contenus, hiérarchisés par leur fréquence sur l'ensemble du corpus. Et lorsque les unités de signification sont de grande taille c'est la meilleure manière de construire une typologie par la méthode de classification hiérarchique descendante. Cependant la généralité du thésaurus peut conduire à des résultats peu concluant ou surprenant. Le résultat obtenu pour les premiers niveaux de l'arborescence n'apporte en général que des évidences, et à un niveau plus fin, on s'aperçoit vite que le domaine abordé est trop spécialisé pour être bien décrit ou que la dénomination des catégories est inappropriée à l'actualité du sujet ou aux spécificités[5] du domaine étudié. Néanmoins et en première analyse l'usage d'un thésaurus général apporte très rapidement le point de vue d'un lecteur extérieur. La construction d'un thésaurus ad hoc évite ces inconvénients mais nécessite une réflexion initiale fondée sur des connaissances préalables et/ou l'exploration lexicale et sémantique du corpus. Elle consiste à définir les idées ou concepts comme on le fait pour élaborer une grille thématique puis à documenter les concepts par la liste des mots qui les définissent en tenant compte du fait qu'ils seront utilisés pour une lecture automatique aux capacités limitées. C'est ce qui a été fait dans un deuxième temps pour analyser globalement les 415 interviews portant sur la fonction RH.

[4] Sphinx utilise le thésaurus Larousse du français développé par Synapse Ingénierie

[5] Par exemple le concept « zoologie » apparaît fréquemment dans des interviews sur la crise Covid. Cela s'explique bien par le mot confinement qui d'un autre point de vue définit les Zoo !

Lien 3 l'importance des thèmes définis par le thésaurus ad'hoc
https://cvip.sphinxonline.net/tiny/v/qMwIJOmaCS

4. De nouvelles approches pour la recherche qualitative

4.1. Fouille de texte

L'opposition classique entre approche inductive et ciblée (Moscarola & ali 2001) passe sous silence le processus de découverte dans lequel se trouve engagé le chercheur grâce à l'usage simultané de la statistique et de l'analyse sémantique qui permettent de focaliser sur tel mot, catégorie thématique ou concept tout en revenant au texte pour fouiller le corpus et en vérifier le sens. L'affichage des résultats sous forme de nuages ou carte de mots, en faisant ressortir graphiquement tel ou tel trait, stimule l'interprétation et conduit le chercheur à naviguer entre conjectures et idées théoriques d'une part, et vérification empirique par la lecture des verbatim d'autre part.

Lien 4 Fouille de texte dans les entretiens sur RH
https://cvip.sphinxonline.net/tiny/v/BeXHej5ny0

En cliquant sur l'une des catégories mise en évidence par la typologie la lecture est focalisée sur l'ensemble des phrases correspondantes qui peuvent être à leur tour fouillées à partir des mots ou expressions clé qui les caractérisent. En outre ce sous ensemble peut à nouveau faire l'objet d'une classification qui permet d'approfondir la typologie de départ et de découvrir plus profondément le corpus.

Lien 5 Fouille de texte dans les documents réseau de chaleur
https://d.ergole.fr/v/SXtI2rtBKV

L'interactivité et la visualisation dynamique permettent de plonger au cœur du corpus découpé en phrases. Tant que le nombre de phrases sélectionnées est important, de nouvelles explorations sont possibles. C'est l'avantage des grand corpus.

4.2. Interprétation, contrôle et partage

La recherche qualitative se distingue par la liberté laissée au chercheur qui s'engage dans un processus créatif potentiellement producteur de connaissance. A lui d'en faire le « storytelling » pour emporter l'adhésion du lecteur. Mais en scientifique il lui appartient aussi d'ancrer son histoire sur le corpus et d'autoriser le débat sur les interprétations qu'il propose. Le recours aux méthodes de l'ADT introduit l'objectivité de l'observation statistique et l'infographie dynamique la possibilité du débat critique fondé sur le partage en ligne des résultats obtenus. Ainsi le lecteur en ayant accès à l'intégralité du corpus peut à son tour faire l'expérience du chercheur pour discuter ses interprétations ou en proposer d'autres.

4.3. Contributions de l'ADT et spécificités de la recherche qualitative

L'ADT donnent la possibilité d'aborder de très grand corpus sans y consacrer tout

le temps que nécessiterait une analyse de contenu traditionnelle menée par simple lecture ou suivant une codification exhaustive. Mais en outre, par l'observation statistique préalable des propriétés et structures du corpus elle objective les rapports du chercheur à son objet d'étude, et par le partage en ligne des analyses elle donne aux lecteurs l'occasion d'expérimenter eux même la découverte du corpus et de discuter les interprétations de l'auteur. Objectivité et débat. Mais ceci passe par une instrumentation couteuse en termes d'apprentissage. Elle représente un obstacle réel qui justifie la constitution d'équipes mixtes associant des chercheurs spécialisés dans la connaissance empirique et théorique du terrain et d'autres maitrisant l'usage des outils et des méthodes. Un autre obstacle tient au risque d'être vite noyé dans l'abondance des résultats et des conjectures qu'ils peuvent susciter. Le choix rapide de quelques pistes d'interprétation fondées sur la clarté des résultats empiriques et des théories qui en démontrent la pertinence pour la connaissance est déterminant. C'est certainement la marque de la recherche qualitative qui dans la complexité des phénomènes qu'elle étudie sait choisir les éléments d'une argumentation susceptible d'emporter la conviction sans se soustraire au débat critique.

Bibliographie

Boughzala Y., Moscarola J. et Hervé M. (2014). Sphinx Quali : un nouvel outil d'analyses textuelles et sémantiques JADT 2014

Dasilva L. (2006). *Thésaurus et systèmes de traitement automatique de la langue,* Documentation et bibliothèque

Drisko JW. et Maschi T. (2016). *Content analysis* Oxford University Press

Dumez H. (2016). *Méthodologie de la recherche qualitative* Vuibert 2016

Hudon M., Deguez D. et Ménillet, D. (2009). *Guide pratique pour l'élaboration d'un thésaurus documentaire* Edition ASTED

Krippendorf K. (2013). *Content Analysis. An introduction to its methodology*. SAGE

Lebart L., Princemin P. et Poudat C. (2021). *Analyse de données textuelles* Presse Universitaire du Quebec

Moscarola J (2018). *Faire parler les données* Edition EMS 2018

Moscarola J. (2018). Data visualisation et infographie dynamique: le logiciel Sphinx. in Kalika Les méthodes de recherche du DBA, Edition EMS

Moscarola J. et Papatsiba V. (2001). Exploration sans a priori ou recherche orientée par un modèle :
Contributions et limites de l'analyse lexicale pour l'étude de corpus documentaires. Actes JADT 2001

Mothe C., Delfosse E. et Bouquet A. M. (2021). L'analyse de données textuelles assistée par ordinateur Les discours sur les réseaux de chaleur, Revue Française de Gestion

Reinert A., (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données, Tome 8, N°2, pp. 187198

Saladaña J. (2013). *The coding manual for qualitative researchers*, SAGE

Co-authorship in Carmen Mola, a case study

Alejandro J. Napolitano Jawerbaum, Patrick Juola

Duquesne University— napolitanojawea@duq.edu
juola@mathcs.duq.edu

Abstract
Spanish authors Agustin Martínez, Jorge Diaz Cortes, and Antonio Santos Mercero are the men behind the female collective pseudonym Carmen Mola. These authors of Spanish Novela Negra (noir fiction), were previously thought to be one female until the three men revealed themselves. The aim of our work is to study what co-authorship and contribution mean for these writers, who have published four books collectively under the same pseudonym. Dividing the books by chapters or pages, we obtained 398 samples, which were tested for similarities and differences with books published independently by each of the authors. Nine analyses were performed, yielding approximately 3663 data points, ranking the authors by distance to each excerpt. We counted the number of times each author ranked in first, second, or third place in their stylistic distance from Mola, then divided these numbers by the total amount of samples for each book, which gave us a normalized number to compare author contribution. Then we calculated the normalized numbers' standard deviation across books, which in our study represents the consistency of contribution, to test whether these numbers were significant.
We found that Santos Mercero was consistently in the first place for three out of the four books, and on average made half of the contributions, whereas Jorge Diaz Cortes was ranked in the second place for the four books. The results for first place across the four books are as follows: Santos Mercero contributed 51% of the total work, and his consistency (as measured by standard deviation) was .0797, Martínez contributed 26% with a standard deviation of .11, and Diaz Cortes contributed 23% with a standard deviation of .066. That being said, about 12% of the data led to a tie between Diaz Cortes and Martínez for second place —something worth discussing since it is open to several interpretations. The fact that Santos Mercero was in first place most consistently can be interpreted in several ways; he could simply be the main or final editor rather than the one who simply wrote the most. Likewise, the 12% tie rate for second place could imply that Santos Mercero is in the middle of the distribution, with a writing style 'between' the two others, and thus ranked closer to Mola as a collective author. Finally, the consistency of contribution for the third place was highest for Martínez, with a tie between Diaz Cortes and Santos Mercero behind him.
The key here is that this provides an interesting case study for a discussion of what stylometry means in a case of co-authorship. These data show that there was a significant difference in percent of contribution towards the works published under the name of Carmen Mola, but this does not mean we can easily explain what form these contributions

took. That will be left up to discussion, as well as the 12% rate of ties for second place.

Keywords: Co-Authorship, Text Analysis, Gender, Stylometry.

1. Introduction

Spanish authors Agustin Martínez, Jorge Diaz Cortes, and Antonio Santos Mercero are the men behind the female collective pseudonym Carmen Mola. These authors of Spanish *Novela Negra* (noir fiction), were previously thought to be one female until the three men revealed themselves. (Graham-Harrison, Emma; Jones, Sam, 2021) The aim of our work is to study what co-authorship and contribution mean for these writers, who have published four books collectively under the same pseudonym.

2. Background

Carmen Mola won several awards, such as the *Premio Planeta de Novela,* which is the richest literary prize in the world. Mola was also considered a major feminist authorial figure, to the point where her books were recommended by the Instituto de la Mujer La Mancha as a feminist reading. A previous (Napolitano Jawerbaum, A. and Juola, P., submitted) work of ours showed that Mola did not present a female authorial persona and was, as a matter of fact, closest to these three authors. This was achieved with a test set of ten male and ten female authors of Novela Negra, using two books by each for a total of forty books. To follow up on this, we decided to investigate what co-authorship looks like in this particular case. The key here is that this provides an interesting case study for a discussion of what stylometry means in a case of co-authorship, more specifically in the Spanish language.

Sarwar, Urailertprasert, et. al. (2020) proposed a framework for authorship attribution (AA) on multi-author documents by attributing different fragments of a co-authored document to different authors given a sample, with both known and unknown co-authors. Following up on their work, in which they point out the low accuracy of stylometric analyses and claim that increasing the number of co-authors adversely affects performance of many methods of AA, we decided to show that stylometry can work in this case where three of the co-authors are known, with the caveat that if there are any other co-authors, they are not known.

Dauber, Overdorf, and Greenstadt (2017) analyzed documents with consistent collaboration (i.e. those where several texts by the same group of co-authors were available), among other instances of collaboration, using non-collaborative training documents in one of their tests, with low-accuracy results using a linear SVM and Naïve Bayes. We expand upon this section of their study by first confirming that the three co-authors can be told apart from other authors of the same genre (see above: Napolitano Jawerbaum, A., and Juola P).

3. Materials and Methods

We used Mola's four books: *La Bestia*, *La Novia Gitana*, *La Red Púrpura*, and *La Nena*, in their original Spanish version. The first three books were divided by chapter due to their length —85, 79, and 65 repectively— whereas *La Nena*, a shorter book, was divided by page by page. This yielded 398 samples, which were tested for similarities and differences with books published independently by each of the authors, namely Agustin Martínez's *La Mala Hierba*, and *Monteperdido*, Jorge Diaz's *Tengo en Mi Todos los Sueños del Mundo* and *Cartas a Palacio*, and finally Santos Mercero's *El Caso de las Japonesas Muertas* and *El Final del Hombre*.

The two books by Martínez totaled 272,129 words, 1,250,834 characters excluding whitespace, and an average word length of 4.6. Those by Santos Mercero had 202,159 words, 912,926 characters, and an average word length of 4.5. Finally, Diaz's books totaled 330,592 words, 1,479,910 characters, and an average word length of 4.47. Meanwhile, all of Mola's works put together had 518,701 words, 2,322,777 characters, and an average word length of 4.46.

Through JGAAP (Juola, 2009), an authorship attribution program, we conducted nine analyses, yielding approximately 3698 data points, ranking the authors by distance to each excerpt. The analyses and distances used were the following:

First, a centroid driver, which computes one centroid —the average relative frequency of events over all documents— using the metric cosine distance, which is a Normalized Dot-Product Nearest Neighbor Classifier.

K-Nearest Neighbors, which yielded more than three places such that we had to ignore results after third place, also with metric cosine distance.

Mahalanobis distance, which is a Generalized Squared Interpoint Distance, meaning a distance between a point and the mean of a distribution; it is unitless and scale-invariant.

Burrows Delta (Burrows, 2002)., which measures the distance between a text of unknown authorship and a corpus with several different authors by comparing how each text diverges from of all of them put together, so this is a good distance measure to ascertain where each author stands in relation to their contributions to Carmen Mola's work.

Finally, we used five analyses from the WEKA package (Frank E. et. al. 2016), namely linear regression, a multilayer and a voted perceptron, naïve Bayes analysis, and SMO. For these analyses, we used the 50 most common and least common words, as well as punctuation. The reason why we chose these methods is that the other methods included in JGAAP either did not fit our methodology, yielded ambiguous results, or returned results that made no sense, such as 0 for most tests and a value of 100% for one of the authors, in direct contradiction with the methods and analyses we used.

After performing these tests, we counted the number of times each author ranked

in first, second, or third place in their stylistic distance from Mola, then divided these numbers by the total amount of samples for each book, which gave us a normalized number to compare author contribution. Then we calculated the normalized numbers' standard deviation across books, which in our study represents the consistency of contribution, to test whether these numbers were significant. A lower standard deviation represents a more consistent rate of contribution.

4. Results

We found significant differences in the consistency of contribution and mean contribution of each author. The *t* test assuming unequal variances yielded *p*-values lower than .05 when comparing the authors for first, second, and third place, with the exception of a tie between Diaz Cortes and Santos Mercero for third place. In these analyses, Antonio Santos Mercero was significantly ahead of the other two for first place, and Diaz Cortes for second place, though there was a 12% tie rate between Diaz Cortes and Martínez for second place. The results for first place across the four books are as follows: Santos Mercero contributed 51% of the total work, and his consistency (standard deviation) was .0797, Martínez contributed 26% with a standard deviation of .11, and Diaz Cortes contributed 23% with a standard deviation of .066.

Table 1. Results of the analyses: Number of times an author was ranked in 1st, 2nd, or 3rd place divided by the total amount of samples, as well as the standard deviation which represents the consistency of contribution.

Book/Pos	Jorge Diaz Cortes	Antonio Santos Mercero	Agustin Martinez
La Bestia	815 total		
1st	0.338650307	0.440490798	0.220858896
2nd	0.436809816	0.261349693	0.358282209
3rd	0.164417178	0.211042945	0.334969325
La Nena	1574 total		
1st	0.184243964	0.529860229	0.285895807
2nd	0.458703939	0.257306226	0.467598475
3rd	0.258576874	0.198221093	0.221092757
La Nov.G	722 total		
1st	0.224376731	0.574792244	0.200831025
2nd	0.501385042	0.398290598	0.433652531
3rd	0.221614227	0.195290859	0.461218837
La Red P	585 total		
1st	0.124786325	0.348717949	0.526495726
2nd	0.620512821	0.398290598	0.343589744
3rd	0.186324786	0.282051282	0.145299145
Average of the above:			
	Jorge Diaz Cortes	Antonio Santos Mercero	Agustin Martinez
1st	0.218014332	0.473465305	0.308520363
2nd	0.504352904	0.328809279	0.40078074
3rd	0.207733266	0.221651545	0.290645016
As a whole: 9 analyses with approximately 398 documents. 3698 results:			
	Jorge Diaz Cortes	Antonio Santos Mercero	Agustin Martinez
1st	0.224952225	0.512421512	0.261807262
2nd	0.493311493	0.228243021	0.404859405
3rd	0.266448266	0.255062945	0.31012831

STDEVS across books; INTERPRETATION: consistency of contribution.			
	Jorge Diaz Cortes	Antonio Santos Mercero	Agustin Martinez
1st place	0.069950739	0.079197742	0.117579832
2nd place	0.063638707	0.074493982	0.046121497
3rd place	0.039963073	0.034452316	0.107088135
AVG each book			
& compiled			
1st place	0.221483278	0.492943409	0.285163813
2nd place	0.498832199	0.27852615	0.402820072
3rd place	0.237090766	0.238357245	0.300386663

Table 2. Results for first place, showing a significant difference in first place between Santos Mercero and Cortes.

t-Tests: Two-Sample Assuming Unequal Variances		
1st place		
Cortes Vs Santos Mer	*Diaz Cortes*	*Santos Mercero*
Mean	0.219170647	0.48016836
Variance	0.004893106	0.006288937
Observations	6	6
Hypothesized Mean D	0	
df	10	
t Stat	-6.045771754	
P(T<=t) two-tail	0.000124304	
t Critical two-tail	2.228138852	

Table 3. Results for first place, showing a significant difference in first place between Santos Mercero and Martínez.

Santos Mercero vs Martinez	*Santos Mercero*	*Martinez*
Mean	0.48016836	0.300734847
Variance	0.006288937	0.013825017
Observations	6	6
Hypothesized Mean Difference	0	
df	9	
t Stat	3.099063392	
P(T<=t) two-tail	0.012741699	
t Critical two-tail	2.262157163	

DC vs Martinez	*Diaz Cortes*	*Martinez*
Mean	0.21917065	0.300734847
Variance	0.00489311	0.013825017
Observations	6	6
Hypothesized Mean	0	
df	8	
t Stat	-1.46030663	
P(T<=t) two-tail	0.18233359	
t Critical two-tail	2.30600414	

Notice that, while there is a significant difference between Santos Mercero and the other two, the difference between Diaz Cortes and Martinez is not significant.

5. Discussion

In the aforementioned previous study, we made a methodological error and included Agustin Martínez in our sample for males and females, and that inclusion made the data look like Mola was overwhelmingly male, more so than the second, proper run without the authors behind Mola. This previous error shows that there are clear elements of the specific authors behind Mola within the work published under that collective pseudonym, which gives credibility to our study in that the data do show collaboration rather than random closeness, especially considering the KNN analyses consistently rate these three authors above all the control samples. These samples included ten female and ten male authors, with two books each, for a total of forty books.

Although it may not seem comparable to have single page samples for one book as opposed to whole chapters in the other books, the data in table 1 show that this does not affect the distribution of our data.

While the data show that there was a significant difference in percent of contribution towards the works published under the name of Carmen Mola, this does not mean we can easily explain what form these contributions took. The fact that Santos Mercero was in first place most consistently can be interpreted in several ways; he could simply be the main or final editor rather than the one who simply wrote the most. Likewise, the 12% tie rate for second place could imply that Santos Mercero is in the middle of the distribution, with a writing style 'between' the two others, and thus ranked closer to Mola as a collective author. Ultimately, the authors themselves hold the answers to these questions. Furthermore, the lack of a significant difference between Diaz Cortes and Martinez may be indicative of a similar level of contribution, but it may also be due to stylistic similarities between the two authors, this will be left for further work.

6. Future Work

Given our results, our future work will center on replication of this study in other cases of co-authorship. We will also investigate the relationship between Mola's work and the stylistic convex hull created by the three authors behind the pseudonym.

References

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–287.

Dauber, E., Overdorf R., and Greenstadt R., "Stylometric Authorship Attribution of

Collaborative Documents." CSCML 2017, DOI: 10.1007/978-3-319-60080-2_9
Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
Graham-Harrison, Emma; Jones, Sam (16 October 2021). "Female Spanish thriller writer Carmen Mola revealed to be three men". *The Guardian*
Juola, P. (2009). "JGAAP: A System for Comparative Evaluation of Authorship Attribution." *Journal of Digital Humanities and Computer Science* 1(1)
Napolitano Jawerbaum, A. and Juola, P. Carmen Mola and stylometry of gender, submitted to DH Unbound, 2022.
Sarwar R., Urailertprasert N., et. al. "CAG: Stylometric Authorship Attribution of Multi-Author Documents Using a Co-Authorship Graph." *IEEE Access*, *8*, 18374–18393. DOI: 10.1109/ACCESS.2020.2967449

From fake news identification to news persuasion level study: the application of a machine learning model with the M.A.D.I.T. methodology

Luisa Orrù, Marco Cuccarini, Christian Moro, Monia Paita, Davide Bassi, Giovanni Da San Martino, Nicolò Navarin, Gian Piero Turchi

University of Padua – luisa.orru@unipd.it christian.moro@unipd.it monia.paita@gmail.com davide.bassi.3@studenti.unipd.it gianpiero.turchi@unipd.it marcocuccarini1@gmail.com dasan@math.unipd.it nnavarin@math.unipd.it

Abstract
A plague that hit the news world in recent years has been the use of fake news to modify the public opinion, Social networks allow people to directly access news without the mediation of any reliable source, so recognizing news as fake is often a difficult task for humans and machine. This paper reconfigure the problem: from a check of the truthfulness of the news content, to the analysis of the text persuasion index, as the way with which this information is introduced to the reader, assuming that fake news is aimed at persuading the reader against the reality sense that intends to convey. The MA.D.I.T. methodology is useful to describe the way by which texts are shaped, overcoming the content or structure analysis level and stressing the study of the Discursive Repertories (DRs). The latter, as discursive modalities of reality sense generation, have been classified into real and fake news categories thanks to the Machine learning application. The dataset consists of 7.387 analysed news. The results highlight different profiles of text generation between the two news groups: the different and typical DRs of the two groups allow to validate the methodological approach used as a good predictor of the news persuasion level.

Keywords: Fake news; Persuasion index; MADIT methodology; Machine learning; Dialogic analysis; Naive Bayes model; Discursive configuration; Reality's sense

1. Introduction

1.1. Fake News: Definitions, Methods and Issues

The phenomenon of fake news is getting more attention, both within the academic community and the citizenry, especially for the harmful effects it can have on community. In the face of this, the scientific community has therefore

worked to provide tools to help citizens identify fake news, informing themselves in an informed manner. Artificial Intelligence and Natural Language Processing (NLP) thus become increasingly important tools to reduce the human workload when dealing with fake news detection (Oshikawa et al., 2020). Currently, several computerised methods are available for detecting fake news, among the main ones: Fact Checking, Style-Based Fake News Detection, Propagation-Based News Detection, Source-Based Fake News Detection (Zhou and Zafarani, 2020; Oshikawa et al., 2020). The plurality of available methods can be traced back to the lack of a specific and shared definition of "fake news". As a matter of fact, currently, the construct of fake news is used to refer to a wide range of phenomena (Andersen and Søe, 2020). At the same time, there are several constructs that are "similar" to fake news, but are distinct from it: satire (Horne and Adali, 2017), people's personal orientations (stance detection; Ferreira and Vlachos, 2016), and clickbait (Chen et al., 2015), could undermine the effectiveness of the algorithm. Moreover, fake news opens some questions like: how do we decide on the truth of a certain content when the news cannot find factual correspondences (Andersen and Søe, 2020)? How much of the analysed content is false, and how to consider the news when it's not completely false (Oshikawa et al., 2020)?

1.2. From Fake News to Persuasion Index through the Propaganda Index

The presented work attempts to abandon the dichotomy "fake-news/real-news" in favour of the creation of an index of persuasion, based on the assumption that fake news is characterised by trying to persuade the reader of a certain news or news's idea (see also Nyilasy, 2019). This perspective is in continuity with Barron-Cedeno and colleagues' (2019) contribution on propaganda (see also Da San Martino et al., 2020). In the attempt to build a model of automated analysis of biased information, the authors move towards an approach that provides an index related to how much the news tries to influence the reader's opinion, through automated analysis of the structure and content of the text: a propaganda index. About the persuasion construct, a theoretical diaspora is traceable, similar to that outlined above, as the work of Miller and Levine (2019) show. This proliferation of theories, in turn, shows critical sides as it can lead (and has led) to conflicting, if not downright contradictory, results on the same topic (Druckman, 2021). O'Keefe (2016) made an attempt to manage this theoretical-methodological problem observing how the manipulation of different persuasive factors affects people's mental states. Therefore, this perspective on persuasion, integrated with the analysis tools made available by the NLP, enables the extraction of the characteristics of the arguments that allow us to effectively change the interlocutor's "position" on a given argument (i.e. "persuasion techniques") (we note the work of Li et al., 2020; Hunter et al., 2019). Following

these recent works on the argumentative structures that characterise persuasive messages, we set ourselves the goal to provide an index measuring the degree of persuasion of the text, discarding the distinction fake vs. real news in favour of a measure that informs the reader about the degree of conviction that the text is exercising on him/her/them. This, through an analysis of the ways in which language is used, enabling the identification of typical modes of persuasive messages. While the persuasion techniques are correlated with specific ill-formulated arguments, we can observe that it is not possible to say how much persuasive a text is (so, to measure it), which is in fact associated with the rhetorical-argumentary architecture of the text itself (namely, how it is discursively structured) and the amount of critical reading competence and attention needed to consume it.

As a theoretical reference to solve this problem, we have chosen Dialogical Science (Turchi, et al., 2021; Pinto et al., 2022; Turchi et al., 2014; Turchi and Orrù, 2014), which, through the formalisation of Natural Language, has given value and rigour to the description of Natural Language's use, formalising the rules of its use. The formalisation of Language has given value to language: defining it, grouping it, and then measuring it (Turchi and Orrù, 2014; Turchi et al., 2014). It is therefore possible to describe how Natural Language is organised, how it generates realities, and also how much it generates. In other words, it's possible to describe the process by which Natural Language is given shape and the weight that language assumes in terms of persuasiveness, allowing the construction of an index. We defined persuasiveness as "*the possibility of a text to get the reader to use discursive modes and contents similar to those of the text itself*". Thank to this approach, we defined a novel index of persuasiveness of a text derived from the Discursive Repertories (DRs - Turchi et al., 2021; Turchi and Orrù, 2014; Turchi et al., 2014), namely specific language use modalities, as defined by the M.A.D.I.T methodology (Turchi et al., 2021; Turchi et al., 2014). There are 24 DRs, each of which identifies one of the possible language use modality that can be traced in a text. The combined analysis of 12 of them, regarding a specific goal (the persuasiveness of a text), allows to state how much a whole text is aligned with that goal. Based upon the specific characteristics and properties of each repertory, has been possible to identify the DRs that could contribute most to making a text persuasive (for the list see tab.1 in the results section), ranking them according to their persuasiveness' potential and creating a persuasion index. In other words, these 12 DRs can be understood as the elements that constitute the DNA of a persuasive news. Their use in conveying a certain piece of information in a text/news increases the possibility that the reader will use modalities and content similar to those carried by the persuasive text, in regard to the topic of the news item. Therefore, the aim of this exploratory experimentation is to observe whether the DRs that, theoretically, should

characterise persuasive texts, characterise the texts of fake news more than the text of the real news. We anticipate that even real news texts could employ DRs related to the construct of persuasion. However, based on the assumption that fake news is particularly characterised by trying to persuade the reader of a certain idea, we expect a greater occurrence of repertories indicating persuasion within fake texts.

2. Methods

2.1. Dataset

To pursue the goal of this exploratory experimentation, we built a dataset containing both fake news and real news using a package of the python library which, after providing the link of the news container site, returns the title, the text and other information of the news itself. The function that downloads the text of the online newspaper is not always perfect, since it uses the html tags to get the parts that are needed from the page. But not all pages have the same structure: if real news usually has quite similar structures, in fake news it often happens that the program also downloads other parts of text, not inherent to the article, since fake news are often blogs or less elaborate pages. We eliminated html tags, spaces and the parts of text that repeated the most. We found an imbalance related to the diversity of topics covered: fake news usually focuses on a certain set of topics, ignoring others. In this case, we tried to do a minimum of preprocessing work, eliminating from real news topics such as sports and news: the first little covered by fake news, the second instead usually conditioned by the events of the period in which the news was written. The dataset is composed of 2776 real news and 4611 fake news; the real news have been extracted from the most famous newspapers in Italy, while the fake ones are extracted from different blogs or newspapers known for the falsehoods of the news.

2.3. Text analysis and Persuasion Index

We built a corpus of human annotated texts and devised a machine learning approach to identify the repertories. Naive Bayes was initially employed, but its low rate in precision (0.35) and recall (0.37) led us to use BERT. With this transformer, we have created a model which manages to divide the inserted text into excerpts, using a sent_tokenise function: a package of the library NLTK, which returns a sentence-tokenized copy of text, using NLTK's recommended sentence tokenizer, after receiving in input the text and the language of the text. Subsequently, the model classifies these excerpts according to the 24 possible categories. The model that is closest to human performance has a precision level of 0.47 (recall=0.43; f1-score=0.43), which can be improved considering that human roles with a basic training have a precision=0.65 (recall=0.63; f1-

score=0.63). Following the construction of the dataset, we used the model to analyse the fake and real news texts, to detect how the chosen model divides the various excerpts and names them, going into details of the distribution of the discursive repertoires for fake and real news. Having for each text, real or fake, the distribution of discursive repertories, we used a model to predict whether a text is a real news or a fake news; this, to provide further support for the theoretical link between persuasion and biased texts, based on the distribution. We then implemented a model called Random Forest, which has a good tolerance to errors and missing data, and also allows us to define the importance that each repertory has had in the classification.

Figure 1 shows the distributions of Repertories according to the text categorization in fake and real news. Generally, a strong presence of characteristic repertories is observable for both the text codifications.

Figure 1. Distribution dataset real and fake news

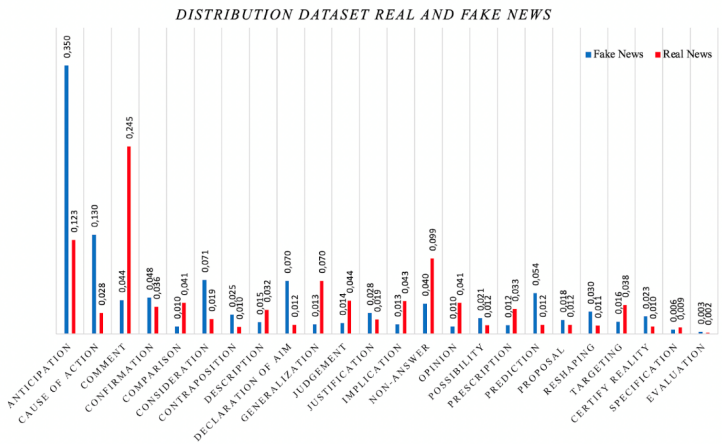


Table 1 adresses the Figure 1's distributions and reports the percentage occurences of the 12 repertories in fake and real text, characterising a persuasive text and the Feature Importance given by the Random Forest (RF). As anticipated in the introduction, the assumption of our research was that we would be able to trace an increased occurrence of persuasion-indicating repertoires with regard to the texts of fake news, as they are marked by a tendency to persuade. Results confirm this hypothesis for seven of the twelve indicated repertoires, specifically for "Anticipation", "Cause of Action", "Declaration of Aims", "Prediction", "Justification", "Certify Reality" and "Evaluation" (The asterisk indicates cases where the percentage of fakes is higher than the real ones). When looking at the results, it is also necessary to consider the imbalance between fake and real news in the dataset, the level of accuracy of the model in relation to the naming task

(0.47) and, in general, the errors behind in the various phases of the program, both by the model that classifies the repertories but also from the one that divides the text. The results in the prediction of a text as real or fake news have been tested with the Random Forest' model: Precision=0.76; Recall=0.76, F1-score=0.76; which can be considered very good results due to the difficulty of the task and the margin of error of the model that predicts the discursive repertories of each text.

Table 1. Distribution of Discursive Repertoires indicators of persuasions between fake and real news

DR	Fake	Real	RF	DR	Fake	Real	RF
Anticipation*	0.35	0.123	0.00937	Targeting	0.016	0.038	0.01921
Cause of Action*	0.13	0.028	0.03254	Judgment	0.014	0.044	0.03176
Declaration of Aim*	0.07	0.012	0.02566	Implication	0.013	0.044	0.03111
Prediction*	0.054	0.012	0.05226	Prescription	0.012	0.033	0.03958
Justification*	0.028	0.019	0.02134	Opinion	0.01	0.041	0.02325
Certify Reality*	0.023	0.01	0.03547	Evaluation*	0.003	0.002	0.02545

4. Conclusions

This work has moved in the direction of providing the community with a tool to deal with fake news and, more generally, with the different news offered by the information platforms available today. Pursuing this objective, in line with the work of Barron-Cedeno and colleagues (2019), a perspective was adopted that attempted to overcome the traditional 'fake-news/real-news' distinction, which is a harbinger of several critical issues. Therefore, starting from the assumption that fake news aims to persuade the reader of something, a persuasion index was constructed to provide the reader with elements to evaluate the bias of a certain piece of news, based on the specific ways in which language is used in a text. The analysis of language use patterns was carried out according to the theoretical-methodological references of Dialogic Science (Turchi and Orrù, 2014) and implemented through an automatic process, based on the BERT transformer. The results obtained from the experiment generally support what has been argued in theory. Specifically, we observed a greater occurrence of Discursive Repertoires, which theoretically should generate persuasion in the reader, in fake news texts, and a greater occurrence of "less persuasive" repertories in real news texts, as was also shown by the results of the Random Forest model. Therefore, it is possible to assert that the elements of the 'DNA' of persuasive news, identified through the Dialogic Science framework, have been matched by experimental data. However, this kind of distribution in support of theory has not been found

in all repertoires that are supposed to promote a process of persuasion. This can be traced back to several factors: the imbalance of the dataset, the error made by the model in processing the text in excerpts, as well as in the naming process. Thus, the findings highlight certain aspects of the model used and the automated textual analysis, which will need to be specified and refined in the future in order to increase the accuracy of the automated analysis. A further future work perspective concerns the construction of the dataset. As a matter, of fact, the need for pre-processing work on the fake news dataset emerged: on the one hand, by balancing the topics covered, i.e. maintaining a similar distribution between the fake news part and the real part; on the other hand, by refining the process of "cleaning" the text from parts that do not relate to the article in question, but are mistakenly downloaded from the Python library.

References

Andersen J. and Søe S.O. (2020). Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication*, 35(2), 126-139.

Barron-Cedeno A., Jaradat I., Da San Martino G. and Nakov P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing and Management*(56), 1849-1864.

Chen Y., Conroy N.J. and Rubin V.L. (2015). Misleading online content: recognizing clickbait as "false news". *Proceedings of the 2015 ACM on workshop on multimodal deception detection* , 15-19.

Da San Martino G., Shaar S., Zhang Y., Yu S., Barrón-Cedeno A. and Nakov P. (2020, July). Prta: A system to support the analysis of propaganda techniques in the news. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 287-293).

Druckman J.N. (2021). A Framework for the Study of Persuasion. *Annual Review of Political Science,* 25.

Ferreira W. and Vlachos A. (2016). Emergent: a novel data-set for stance classification. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human lan- guage technologies,* 1163-1168.

Horne B.D. and Adali S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *The 2nd International Workshop on News and Public Opinion at ICWSM*.

Hunter A., Chalaguine L., Czernuszenko T., Hadoux E. and Polberg S. (2019). Towards computational persuasion via natural language argumentation dialogues. *In Joint German/Austrian Conference on Artificial Intelligence* (Künstliche Intelligenz) (pp. 18-33). Springer, Cham.

Institute for Propaganda Analysis (1938). How to detect propaganda. *Propaganda analysis*. volume 1 of the publications of the institute for propaganda analysis, 201-218.

Jowett G.S. and O'Donnell V. (2018). Propaganda and Persuasion. *Sage Publications*.

Li J., Durmus E. and Cardie C. (2020). Exploring the Role of Argument Structure in Online Debate Persuasion. *arXiv*.

Miller M.D. and Levine T.R. (2019). Persuasion. In D. W. Stacks, M. B. Salwen, & K. C. Eichhorn, *An Integrated Approach to Communication Theory and Research* (p. 261 - 277). New York: Routledge.
Nyilasy, G. (2019). Fake news: When the dark side of persuasion takes over. *International Journal of Advertising,* 38(2), 336-342.
O'Keefe D.J. (2016). Persuasion. *Theory and Research*. London: Sage.
Oshikawa R., Qian J. and Wang W.Y. (2020). A Survey on Natural Language Processing for Fake News Detection. *LREC, 6086-6093.*
Pinto E., Alfieri R., Orrù L., Dalla Riva M.S. and Turchi G.P. (2022). Forward to a methodological proposal to support cancer patients: the Dialogic contribution for the precision care. *Medical Oncology* 39(5):75. doi: 10.1007/s12032-021-01644-1.
Turchi G.P., Dalla Riva M.S., Ciloni C., Moro C. and Orrù L. (2021). The Interactive Management of the SARS-CoV-2 Virus: The Social Cohesion Index, a Methodological-Operational Proposal. *Frontiers in Psychology*, 12.
Turchi G.P. and Orrù L. (2014). Metodologia per l'analisi dei dati informatizzati testuali: fondamenti di teoria della misura per la scienza dialogica. *Edises*.
Turchi G.P., Romanelli M., Bonazza F. and Girardi A. (2014). Discursive Configuration. In Teo T. (Ed.) *Encyclopedia of Critical Psychology.* Springer: New York.
Zhou X. and Zafarani R. (2020). ASurvey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Survey*, 53(5), 40.

Biographical change and Machine Learning with MADIT: a predictive model proposal for data and intervention analysis

Luisa Orrù, Jessica Neri, Marco Cuccarini, Antonio Iudici, Nicolò Navarin, Gian Piero Turchi

University of Padua – luisa.orru@unipd.it
jessicaneri24@gmail.com marcocuccarini1@gmail.com
antonio.iudici@studenti.unipd.it nnavarin@math.unipd.it
gianpiero.turchi@unipd.it

Abstract
This work is within the discipline of clinical psychology, which is characterized by goals of change and health promotion. On this topic, it is right by virtue of the potential and real impact of questions for biographical change, that Machine Learning (ML) technology, applied through the M.A.D.I.T. methodology, finds its place. The output data of the analysis, made through BERT model, allow to isolate and analyze different sense discursive configurations, through a predictive model that denominates them against the question. The dataset used is made by 14.567 excerpts, totaling about 303.316 graphical forms. The results open to considerations about the usage value of questions and their implications, as well as the "predictivity" degree of change against the intervention aims, developing a precise and rigorous analysis supported by ML techniques for their analysis and measurement.

Keywords: machine learning, MADIT methodology, NPL, psychological intervention, psychotherapy.

1. Introduction

Although psychological and psychotherapeutic interventions generate effects that are positively judged in the literature, it is possible to mention some critical aspects that may reduce their effectiveness in relation to their objectives of change and health promotion. For example, negative outcomes reported by users are mentioned attributed to user characteristics, the role of the counsellor and the interactive modes used (Saxon, et al., 2017), but also critical aspects related to the difficulty of monitoring and anticipating the progress of the change path (Aafjes-van Doorn et al., 2021). In recent years, Artificial Neural Networks

(NNs) have been applied to many different domains, often improving the state-of-the-art in terms of predictive performance. In clinical psychology, preliminary work using natural language processing (NLP) and machine learning (ML) is aimed at: classifying intervention models (Dwyer et al., 2018); identifying behavioral descriptors for monitoring specific clinical situations (Singla et al., 2018); and generating fidelity codes (Tanana et al., 2016). Such research is based on thematic analyses of the content, and not of the interactive process between participants in the clinical context. How can the level of precision and rigor of psychological models of intervention be increased using ML? Here we find the shift needed to answer this question: from the analysis of the content of the text to the study of the question that generates it.

2. The natural language and its analysis: the MADIT methodology

This research refers to the Dialogical Science as theoretical and methodological reference (Turchi et al, 2014 ; Turchi et al. 2021; Pinto et al., 2022) that places itself within the Narrativistic Paradigm. Its object of investigation is in fact natural language, as a tool for generating configurations of realities of meaning. These can be described starting from the rhetorical-argumentative links building them, and not only from the content they convey. By referring to the MADIT methodology (Methodology of Analysis of Textual Computer Data), declination of Dialogical Science, researchers can analyze discursive processes which, by interacting, generate configurations of realities of sense. Dialogical science has made possible to formalize 24 rules of language use, called Discursive Repertories (DRs - Turchi et al., 2021). They are organized within the semi-random Periodic Table of Discursive Repertories, according to their properties or rules of reality generation, defined a priori. The DR are divided into three types. The generative type is a narrative that brings variability to the dialogical process. The maintenance type tends to build a "stable" narrative sense of reality, closed to possibilities of change. The hybrid one can have a maintenance or generative meaning, depending on how the repertories interact in the configuration. Every discursive production belongs to one of these three types, depending on the capacity of triggering a transformative impact (Turchi et al., 2021). Thus, in the clinical field, what makes an intervention potentially effective concerns the possibility of analyzing the generative capacity that the used narrations imprint on the reality of meaning generated in therapy. Hence, the management of uncertainty in discursive configurations becomes a central issue, and a requirement for the counsellor and clinical practice. The MADIT methodology thus enables us to observe the movement of generation of discursive configurations, from the contents of natural language (symbolic units) linked together through the rhetorical-argumentative rules of discursive repertories.

2.1. The role of clinical question and its uses for change

The starting point of any clinical intervention is the definition of an objective, aimed at constructing and generating the text, object of the work. Therefore, the question becomes a fundamental strategy that allows the continuous generation of further text productions, i.e. DRs (Turchi et al., 2021), depending on the sense of reality they generate. The researchers then distinguished the questions into four types, according to the purpose to which they tend, and thus the use value they can generate, so questions aimed at: description, which collect and expand the user's text, allowing different possibilities to be generated; evaluation, which examine the text offered and specify it, expanding the panorama of possibilities; change, which move the configuration in other directions; maintaining, which stabilize the configuration of reality generated during the intervention.

3. Aim and method

About the questions with the aim of description and evaluation, we describe a portion of research carried out at the University of Padua. It aims to investigate the learning and recognition of natural language usage rules by Machine Learning, applied through the MADIT text analysis methodology. In particular, in this line of research, the impact of the type of question on the generation of the discursive configuration was studied.

3.1. Data collection and analysis

A total of 14,567 excerpts[1] collected during past social work research were processed. The dataset has 303,316 graphic forms (14,507 RDs). For the evaluation of DRs naming accuracy, specific metrics were implemented for all machine learning models tested: *Precision*, the number of correctly labeled items out of all items that were labeled (correctly or not) with that class of DR; *Recall*, the number of correctly labeled items compared to the total number of items that actually belong to that DR class; *F1-score*, harmonic mean between Precision and Recall; *Accuracy*, the number of correct predictions out of all predictions made. Models tested include: the *Naive Bayes*, as a set of ML learning algorithms that exploit Bayes' theorem; the *Bidirectional Encoder Representations from Transformers*, an alternative to the most used transformers as it analyzes the text from left to right and vice versa. As shown in Table 1, we chose to use BERT instead of Naive Bayes approach because of the better results offered.

[1] Text excerpt means the minimum portion of text to which the named DR corresponds.

Table 1. BERT model structure

Model	**Pretrained Weights**	**Batch Size**	**Learning Rate**	**Freeze to**
BertClfier	bert-base-italian-xxl-uncased	16	1e -05	1
AdamW Eps	**Max Epochs**	**Patience Epochs**	**Embed Dim**	**Activation Function**
0.0001	200	20	768	ReLU

To confirm what was hypothesized about the value of the question as a tool for generating change in clinical settings, we compared text naming across ML models in two different experimental situations, where the following were offered to the ML model: 1) only the text of the model response, without the question information; 2) the question information and the text of the responses. The question information to the model was passed in two steps: by passing the entire question string, and then by passing the information of the existence of two different question types based on the objective.

4. Results

In Table 2, we can see how the absence of demand slightly affects classification performance. The values of Precision tend to decrease, even in a very important way (with the information of the presence of the question the model has lower performances), with some exceptions (Justification, Confirmation and Anticipation, Possibility and Prescription). A limit case is constituted by the Consideration DR, for which the produced values pass from 0.09 with the use of the question to 0 without its use. This behavior of the model is also evident for the value that corresponds to the Recall, going from a 0.31 to 0. The variations in performance, which in the two experimental cases range from 0.01 (a value also attributable to randomness affecting neural network training) to 0.1 (not attributable to randomness), indicate that the mere information of the presence of an unspecified variable, such as demand, constitutes a naming change element for some DRs. In previous experimental tests, neither the full text of the question nor the reference DR of the question had performed incrementally in the model. In conclusion, informing about the presence of the demand in the naming dataset brings an advantage (in some cases, relevant) in the quality of the naming. Anyway, it affects much less than it does in the case of a human denominator. This finding allows to develop the specification of the question information to be passed to the model. An interesting aspect is that the denomination allows for distribution into different DRs, e.g., with respect to the evaluation question takes over the Possibility DR.

Table 2. Classification performance
(A: with the question; B: without the question)

Discursive Repertories	PrecA	PrecB	RecA	RecB	F1A	F1B
Anticipation	0.06	**0.07**	**0.13**	0.07	**0.08**	0.07
Cause of Action	**0.40**	0.32	**0.48**	0.44	**0.43**	0.37
Comment	**0.26**	0.20	**0.22**	0.20	**0.23**	0.20
Confirmation	0.25	**0.27**	**0.35**	0.31	**0.29**	**0.29**
Consideration	**0.09**	0.00	**0.31**	0.00	**0.14**	0.00
Contraposition	**0.48**	0.43	**0.64**	0.63	**0.55**	0.51
Description	**0.68**	0.64	**0.46**	0.45	**0.55**	0.53
Declaration of Aim	**0.30**	0.26	**0.60**	0.53	**0.40**	0.35
Generalization	**0.23**	0.16	**0.17**	0.11	**0.19**	0.13
Judgment	**0.63**	0.56	**0.52**	**0.52**	**0.57**	0.54
Justification	0.27	**0.28**	**0.42**	0.36	**0.33**	0.31
Implication	**0.31**	0.30	**0.41**	0.34	**0.35**	0.32
Non answer	**0.42**	0.38	**0.62**	0.61	**0.71**	0.47
Opinion	**0.52**	0.46	**0.67**	**0.67**	**0.59**	0.54
Possibility	**0.40**	**0.40**	**0.43**	0.39	**0.42**	0.39
Prescription	**0.37**	**0.37**	**0.74**	0.67	**0.49**	0.48
Prediction	**0.26**	0.24	**0.52**	0.40	**0.34**	0.30
Proposal	**0.28**	0.23	0.52	**0.58**	**0.36**	0.14
Reshaping	**0.31**	0.06	**0.57**	0.14	**0.40**	0.08
Targeting	**0.37**	0.29	0.51	**0.57**	**0.43**	0.39
Certify Reality	**0.55**	0.50	**0.25**	0.16	**0.34**	0.24
Specification	**0.33**	0.30	0.39	0.32	0.36	0.31
Evaluation	0.29	**0.30**	**0.34**	**0.34**	0.31	**0.32**
Accuracy	-	-	-	-	**0.43**	0.39
Macro avg	**0.35**	0.30	**0.45**	0.39	**0.38**	0.32
Weighted avg	**0.47**	0.43	**0.43**	0.39	**0.43**	0.38

4.1. Analysis with the question information

Considering the above results, we wondered what information contained by the question is used by human intelligence, in order to produce a text that differs when faced with questions that are different from each other. Therefore, we provide the dataset with the additional information regarding the objective of the question (describe or evaluate) by analyzing the distribution of DR emerging from different types of questions. The original dataset was modified by splitting the responses against questions aimed at description (total, 12,737) and those aimed at evaluation (total, 324). The data were normalized in order to perform a comparison. The results show that the two distributions have a set of DR that remain identical between the two groups, while not having proportional values. The differences between the distributions of the two groups of questions lie in the central part of the distribution: for the group of responses pertaining to the

question with description value, the DRs Anticipation, Reshaping and Confirmation are present exclusively. Moreover, two significant differences emerge in the two distributions regarding the Opinion and (threshold-valued) Comment DRs: for the question aimed at evaluating, these have values within the range of maximally present DRs (within 0.08 probability), while for the question aimed at describing they are beyond this threshold, distributing themselves as the remaining less incident DR.
In theoretical terms, this empirical finding would confirm the differing valence of the two questions, generating response profiles that, similar in high frequencies, introduce different DRs as possible.

Table 3. DRs' distribution according to description and evaluation questions

Description	**Occurence**	**Probability**	**Evaluation**	**Occurence**	**Probability**
Certify Reality	2213	0.17	Certify Reality	75	0.2
Judgment	2173	0.17	Judgment	69	0.18
Description	1885	0.15	Description	52	0.14
Opinion	687	0.05	Opinion	44	0.12
Comment	606	0.05	Comment	31	0.08
Evaluation	579	0.05	Specification	16	0.04
Specification	538	0.04	Evaluation	13	0.03
Generalization	504	0.04	Generalization	12	0.03
Contraposition	406	0.03	Contraposition	11	0.03
Implication	391	0.03	Prescription	11	0.03
Justification	343	0.03	Possibility	7	0.02
Cause of Action	315	0.02	Proposal	6	0.02
Prediction	278	0.02	Justification	5	0.01
Possibility	269	0.02	Cause of Action	4	0.01
Declaration of Aim	267	0.02	Prediction	4	0.01
Non-Answer	251	0.02	Implication	3	0.01
Prescription	208	0.02	Non-Answer	3	0.01
Confirmation	186	0.01	Consideration	2	0.01
Targeting	173	0.01	Declaration of Aim	2	0.01
Proposal	96	0.01	Targeting	2	0.01
Anticipation	47	0.003			
Consideration	41	0.002			
Reshaping	2	0.0001			

5. Conclusion

Information regarding the existence of the question does not affect the accuracy of the text naming of the responses. Therefore, knowing that a question exists and reading the question through the referred DR is not useful information in itself. So, the authors wondered what information the human expertise was using that they had not yet offered to the ML model, thus what value they could offer to the machine to increase the accuracy of its naming. The exploration focused on the type of question and its typical DR distribution. This has allowed to validate the theoretical assertion according to which questions of different typologies can generate different configurations, anticipable and therefore governable by the expert for the generation of the most useful configuration for the patient. These first experimental data could lead to future insights, which may allow the scientific community and experts in the field to reorganize their research and their interventions in accordance with the direction that, with greater possibility, they intend to give to the change. Future developments include repeating the experiment to have a confidence interval on the calculated metrics and specifying hypothesis testing to compare the cases.

References

Aafjes-van Doorn K., Kamsteeg C., Bate J. and Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, (31*)*1, 92-116. DOI: 10.1080/10503307.2020.1808729

Dwyer D.B., Falkai P. and Koutsouleris N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology,* 14(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Pinto E., Alfieri R., Orrù L., Dalla Riva M.S. and Turchi G.P. (2022). Forward to a methodological proposal to support cancer patients: the dialogics' contribution for the precision care. *Medical Oncology*, 39(5), 1-7.

Singla K., Chen Z., Flemotomos N., Gibson J., Can D., Atkins D. and Narayanan S. (2018). Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. *Interspeech,* 3413–3417. https://doi.org/10.21437/Interspeech.2018-2551

Tanana M., Hallgren K.A., Imel Z.E., Atkins D.C. and Srikumar V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment,* 65, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006

Turchi G.P., Dalla Riva M.S., Ciloni C., Moro C. and Orrù L. (2021). The Interactive Management of the SARS-CoV-2 Virus: The Social Cohesion Index, a Methodological-Operational Proposal. *Frontiers in Psychology*, 12.

Turchi G.P., Romanelli M., Bonazza F. and Girardi A. Discursive configuration. *Encyclopedia of Critical Psychology*, 2014, 457-463. https://doi.org/10.1007/978-1-4614-5583-7_569

Similarities and disparities among French modal verbs through co-occurrence correspondence analysis

Aylin Pamuksaç

University of Neuchâtel – aylin.pamuksac@unine.ch

Abstract
The French modal verbs *devoir*, *pouvoir*, *falloir* and *vouloir* are rarely considered in one same study although they can all express the same main modalities which are the epistemic and the deontic values. In this paper, we want to show that despite the fact they are polysemous and seem to be interchangeable in some context, they are quite different in their use. To do so, we show how close or distant they are through correspondence analysis (CA). This study is innovative because this methodology will be used to determine proximities and distances between words and not between parts of the text, as is usually the case. The CAs are constructed on co-occurrences that modal verbs have with their verbal complementation and the subject pronouns that govern them. Not only do we compare two parameters to see if the properties of the modal verbs are stable but, we also conduct this research in four different corpora, two journalistic, and two encyclopedic to show regular patterns between genres.

Keywords: modality, French modal verbs, co-occurrents, specificity, correspondence analysis, corpora studies.

1. Introduction

Studies on modal verbs in French are numerous but often focus on only one or two verbs and mainly on *devoir* and *pouvoir*. For example, they compare the deontic and the epistemic values of these two verbs (Barbet 2012; de Saussure 2014; Sueur 1977, 1979; Vetters 2004, 2012). However, we know that *falloir* and *vouloir* are also considered modal verbs in French and can also express the two modalities mentioned:

Deontic modality:

1a. Tu **dois** ranger ta chambre.
1b. **Peux**-tu ranger ta chambre ?
1c. Il **faut** ranger ta chambre.
1d. Je **veux** que tu ranges ta chambre.

Epistemic modality:

2a. Il **doit** être bête pour croire que la terre est plate.
2b. Il se **peut** qu'il soit bête pour croire que la terre est plate.
2c. Il **faut** être bête pour croire que la terre est plate.
In another context and in some French-speaking regions, we can also hear[1] this epistemic value with *vouloir*:
3. Il **veut** pleuvoir ce soir.
As we can see from examples 1 to 3, not only *devoir* and *pouvoir*, but also *falloir* and *vouloir* are polysemous and fit into different modal categories. These classifications, although accepted and still widely studied in modality studies, do not allow us to dissociate the linguistic forms from each other. These qualitative studies therefore do not grant us to determine whether a verb form is preferred according to the modality expressed.
In this article, we want to disregard these modal categories in order to study how French modal verbs are similar or different in their actual functioning. To do so, we will study their associations with the 50 most frequent verbs of our corpora but also with their subject pronouns through correspondence analysis (now CA). These two variables were selected for our study because modal verbs are among the only verbs that must be associated with other verbs to have a meaning (Abeillé et al. 2021: 131) and because all verbs do have a subject in French (Lazard 2009: 152-153). Thus, these elements, because there are common for the verbs studied here, are particularly relevant to considerate.
We will conduct this research in four different corpora, all of which are informative. This variety of corpora will allow us to see if regularities of properties occur for French modal verbs across sub-genres of informative text but also across time as we are working on three contemporary corpora and one from the 18th century.

2. Methodology

Two journalistic corpora, one national (*Le Monde* 2010, 17895009 tokens, now LM) and one regional (*L'Est Républicain* 2010, 25825949 tokens, now ER) as well as two encyclopedic corpora, one contemporary (*Wikipedia* 2019, 18715455 tokens, now WIKI) and one from the 18th century (*L'Encyclopédie de Diderot et d'Alembert*, 24239478 tokens, now EDDA) were used for this study[2]. All these

[1] We use here the verb to hear because this kind of construction appears mostly in the spoken language.
[2] All the corpora, with the exception of EDDA, were constructed by ourselves or by members of the team in which we are employed. For EDDA, we thank the members of the DISCO project, and more particularly D. Vigier and A. Brenon.

corpora were explored in the TXM software, developed at the ENS of Lyon (Heiden et al. 2010) to count cooccurrences between our four modal verbs and their verbal complements and their subject pronouns.

2.1. The CA methodology

Developed by Benzécri (1973), the goals of CAs are to summarize a large amount of data of a table in a graphical visualization form in order to bring out contrasts (Poudat & Landragin 2017: 104) but also proximities between the various represented elements (Leblanc 2015: 44). CA is not a new tool in the field of textual data analysis. However, it has rarely been used to study co-occurrences. Usually, this method is used to bring out features of a speech[3]. In our research, we will use this method, not to make sense of parts of a text, but to study the meaning of closely related linguistic forms, in this case French modal verbs, to see whether differences or commonalities emerge in their actual use.

To do so, we produced the CAs in the software Hyperbase (Brunet 2011). TXM can provide CAs but these do not concern co-occurrences but frequencies of lemmas, words or grammatical categories such as verbs, nouns, adjectives, etc. These CAs would show the distribution of these different elements in relation to the partitions of a corpus and thus give us information about the construction of the text rather than the functioning of the words[4].

For our article, because we want to see the proximities or distance between verbs, we counted the number of co-occurrence for the four modal verbs with each verb from our list in every corpus. Our counting was made on TXM and rather manual. Indeed, we went through the KWIC tool of the software to see if the co-occurring verb was really modalised by one of our modals. For example, we omitted uses such as:

4. *Corneille* ***dût*** *même* ***se résoudre à demander*** *l'aide à trois auteurs* (LM)

where *demander* is linked to *se résoudre* and not the verb *devoir*.

[3] See for example Mayaffre et al. (2019) for a study on the similarities and differences between speeches by French presidents.

[4] For example, TXM could have provided us with the distribution of the various verbs we are studying here across the different EDDA authors or whether one uses more adjectives than another but not how close or distant two verbs are in their actual uses.

3. Discussion on the results

3.1 The co-occurrences between the modal verbs and their verbal complements[5]

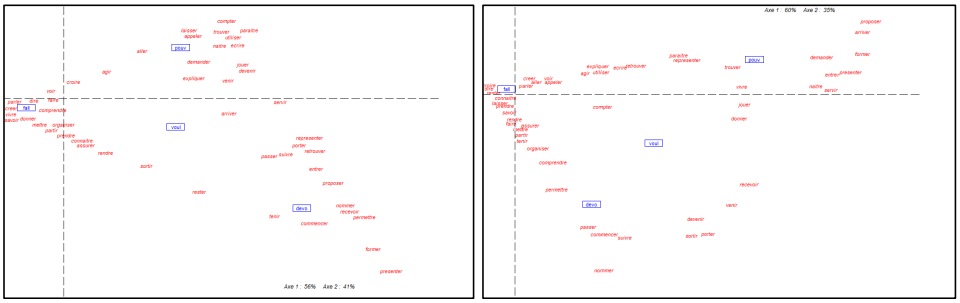


Figure1. CA of co-occurrences between the four French modal verbs and the 50 most frequent verbs of our corpora. Left LM, right WIKI.

Falloir shows a very stable behavior as in all corpora it is found in the center of the two axes and surrounded by verbs such as *croire*, *parler* and *savoir*. *Vouloir* also adopts a regular behavior in all corpora. Indeed, there seems to be a "void" around it, except in ER, where it is surrounded by *expliquer*, *nommer*, *arriver*, *naître* and *commencer*.

Another regularity is that in all CAs, two scales seem to stand out from *falloir*. In fact, some co-occurrents make the link between *falloir* and *pouvoir* towards the top of the graph and others move under, to *devoir*. *Falloir* thus seems to be at the intersection between *devoir* and *pouvoir*.

Indeed, while the vertical axis is similar in the four corpora, this is not the case for the horizontal axis. From the left, we have three possible patterns:

I) ***Falloir*** ⟶ **à** ***pouvoir-vouloir*** ⟶ ***devoir*** in the two press corpora: here, we can see that the verbs of necessity are very far from each other. This information is even more important as the horizontal axis contains more information than the vertical axis, so this distinction is all the more important to note. *Pouvoir* and *vouloir*, on the other hand, share more or less the same position on this axis, point that has not been found in studies on modality yet. The qualitative studies on modality did not allow us to identify such proximity between these two verbs because they were never studied together, as we said in our introduction, most

[5] For reasons of space, we have included here only the CAs of the LM and WIKI corpora. These two CAs as well as the two missing CAs are visible in optimal quality under: https://drive.switch.ch/index.php/s/vil2d3FnYX2hzdq

of the time, *pouvoir* is compared to *devoir*. Our methodology demonstrates that this point deserves to be analyzed, or at least considered, in qualitative research.
II) ***Falloir*** ⟶ ***devoir*** ⟶ ***vouloir*** ⟶ ***pouvoir*** in WIKI: this scale seems much more predictable than the previous one if we consider the qualitative research. Indeed, *falloir* and *devoir* are both verbs of strong necessity, mostly used to express an obligation. This proximity between *devoir* and *falloir* in WIKI can be explained by the fact that *devoir* expresses more often the deontic modality in this corpus than an epistemic or quasi-certainty modality that we were able to identify in a previous study (Rossari et al. 2021). Thus, we should see in the other corpora whether the distance between *devoir* and *falloir* is not simply because *devoir* does not express deontic modality as often as one might imagine.
III) ***Falloir*** ⟶ ***vouloir*** ⟶ ***devoir*** ⟶ ***pouvoir*** in EDDA: in the 18th century corpus, another pattern seems to emerge. Volition seems to link two modal verbs expressing, *a priori*, necessity and *pouvoir* and *devoir* seem closer than in the other corpora. In addition, many verbs are found between *pouvoir* and *devoir* on the vertical axis which shows a certain sharing of co-occurrents between these two verbs.

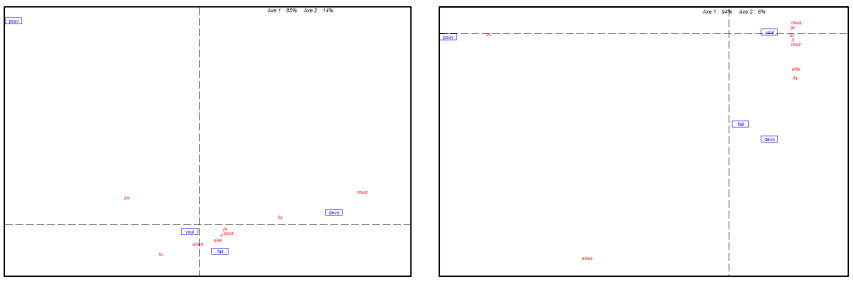


Figure 2. CAs of co-occurrences between the four French modal verbs and their subject pronouns in LM (left) and WIKI (right)

3.2 The co-occurrences between the modal verbs and their subject pronouns

We wanted to see if the subject pronouns of these modal verbs can also make appear some similarities or differences that the classic modal classification does not make come up. We did the CA here only for two corpora, one in each genre, to see if the differences and similarities already denoted before with the verbal complementation emerge again or not.
First of all, it could be surprising to see *falloir* in this kind of results because it is an impersonal verb, so it is always used with the third person singular male, i.e. *il* in French. But our primary goal was to consider all modal verbs for our study, regardless of their syntactic specificity or qualitative analyses about them. This is why we have taken *falloir* into account but counting the occurrences of

the complement pronouns that appear before the verb as in
5. Il **me** fallait gagner plus de temps. (LM)
In which *me* was counted like an occurrence of *je*. We thus considered all the cases where *falloir* was preceded by a pronoun in the dative, and even sorted the occurrences of *lui* and *leur* to see if they represented feminine or masculine nouns in order to count the number of occurrences of *elle* and *il*. Incidentally, we have therefore omitted *il* pronoun from our calculations[6].
In the journalistic corpus, *vouloir* is very central and demonstrates some proximity with first-, second- or third-person pronouns. In contrary, in WIKI, this verb stands out from the others by attracting mainly first or second person pronouns, which seems, in this corpus, to be a specific feature of *vouloir*, which is on the extreme right of axis 1 with all these pronouns.
On the other extreme, in both corpora, *pouvoir* seems to have a behavior that stands out strongly from the others. The pronoun which put *pouvoir* in such position is without doubt *on*. This phenomenon could explain the proximity between this verb and *falloir* we already identified in the press corpora with the verb complements in §3.1. This closeness also appeared to us in a previous study, in which we only considered one semantic category, the *dicendi* and *putandi* verbs as complements to the modal verbs in our CAs (Pamuksaç 2022). *Pouvoir*, with *on*, could denote some "abstract effect", already noted by Larsen (in Fløttum et al. 2007: 22-23) which put it closer to *falloir*, an impersonal verb.
Finally, we also note that in WIKI, again, *falloir* and *devoir* are closer in their associations with subjects than in LM. This proximity between these two verbs seems to be a stable aspect of this corpus, whatever the parameters considered.

4. Conclusion

In this study, we wanted to compare all four French modal verbs (*pouvoir*, *devoir*, *falloir* and *vouloir*) in a quantitative study to see if there were similarities or differences between them on various parameters. Indeed, as explained in the introduction, modal categories in qualitative studies allow us to see that modal verbs are all polysemous but not if the closeness of meaning between them. For this purpose, we used the CA representation to see whether in their actual use the modal verbs tended to resemble each other or to denote certain tendencies. It appeared that by studying their verbal complementation, several scales were

[6] In our research, the study of pronouns is still in its infancy. For example, we do not know what to make of cases without a pronoun in the dative with *falloir*. Perhaps we can consider them as an impersonal on, and thus count these occurrences as the pronoun on. For the moment, we have considered that this pronoun does not appear with *falloir*.

emerging, ranging from a modality denoting a strong obligation to a possibility, i.e. a much weaker necessity. Even if certain regularities seemed to emerge, differences between corpora, and therefore perhaps between genres and periods, were apparent. But, we also denoted that some properties of these modal verbs are stable, despite the parameters we study: the proximity between *falloir* and *devoir* has been made salient in WIKI, both with verbal complementation and with subject pronouns but not in other corpora for example.

Our methodology can thus fill a gap in the study of the polysemy of French modal verbs. For now, we noticed some proximity or some distance between them with two criteria studied in four corpora, but we could deepen this study and determine clearer proximities or disparities between modal verbs, if, in the future, we label them with their values. Thus, some complement verbs might appear more often with *falloir* and *devoir* when both express necessity, and others would be closer to *pouvoir* and *devoir* when they denote an epistemic modality. In that case, we could therefore show that the modal verbs are similar or note in their uses according to their values.

References

Abeillé A., Koenig J. P., Godard D. and Bonami O. (2021). Qu'est-ce qu'un verbe ? In A. Abeillé, D. Godard, A. Delaveau, & A. Gautier (Eds.), *La Grande Grammaire du Français* (Vol. 1, pp. 127-147). Actes Sud.

Barbet C. (2012). Devoir et pouvoir, des marqueurs modaux ou évidentiels? *Langue française*, 1: 49-63.

Benzécri J. P. (1973). *L'analyse des données*. Dunod.

Brunet E. (2011). *Hyperbase, Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels, Manuel de Référence*. Université de Nice.

de Saussure L. (2014). Verbes modaux et enrichissement pragmatique. *Langages*, 193 (1): 113-126.

Fløttum K., Jonasson K. and Norén C. (2007). *On : pronom à facettes*. De Boeck/Duculot.

Lazard G. (2009). Qu'est-ce qu'un sujet? *La linguistique*, 45 (1): 151-158.

Leblanc J. M. (2015). Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie. *Nouvelles perspectives en sciences sociales*, 11 (1): 25-63.

Mayaffre D., Pincemin B. and Poudat C. (2019). Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours. *Langue française*, 3: 101-115.

Pamuksaç A. (2022). Les verbes modaux du français : ressemblances et dissociations à l'aide de leurs associations avec les verbes *dicendi* et *putandi*. *8ème Congrès Mondial de Linguistique Française 2022 (CMLF2022)*, https://doi.org/10.1051/shsconf/202213811014

Poudat C. and Landragin, F. (2017). *Explorer un corpus textuel : Méthodes, pratiques, outils (Champs linguistiques. Recherches).* De Boeck Supérieur.

Rossari C., Montrichard C., Pamuksaç A., Ricci C. and Sanvido L. (10 novembre 2021). Connecteurs et modalités : des traces pour identifier les choix rhétoriques dans un discours. *Colloque international ARGAGE 2021 (Argumentation & Langage)*, 10-12 novembre 2021.
Sueur J. P. (1977). À propos des restrictions de sélection : les infinitifs devoir et pouvoir. *Lingvisticae investigationes*, 1 (2): 375-409.
Sueur J. P. (1979). Une analyse sémantique des verbes devoir et pouvoir. *Le Français Moderne Paris*, 47 (2): 97-120.
Vetters C. (2004). Les verbes modaux pouvoir et devoir en français. *Revue belge de philologie et d'histoire*, 82 (3): 657-671.
Vetters C. (2012). Modalité et évidentialité dans pouvoir et devoir : typologie et discussions. *Langue française*, 1: 31-47.

Le temps dans le discours, expérimentation d'un protocole d'observation des caractéristiques temporelles d'un corpus d'avis de salariés

Daniel Pélissier, Jérôme Bousquié

Université Toulouse Capitole - IUT Rodez - IDETCOM -
daniel.pelissier@iut-rodez.fr jerome.bousquie@iut-rodez.fr

Abstract
Time and language are closely linked. Lexicometry has developed methods for understanding speech over time. This article proposes an orthogonal problematic by studying time in language. The experiment concerns a structured method for analyzing time in discourse from a corpus of massive data of employee reviews. The protocol used makes it possible to approach meaning through temporal specificities. The observed limits of this exploratory approach justify further research on this original question.

Keywords: time, experimentation, review, protocol

Résumé
Temps et langage sont étroitement liés. La lexicométrie a développé des méthodes pour comprendre les discours dans le temps. Cet article propose une approche complémentaire en étudiant le temps dans le langage. L'expérimentation porte sur une méthode structurée d'analyse du temps dans le discours à partir d'un corpus de données massives d'avis de salariés. Le protocole utilisé permet d'aborder le sens par les spécificités temporelles. Les limites constatées de cette approche exploratoire justifient une poursuite des recherches sur ce questionnement original.

Mots clés: temps, expérimentation, avis, protocole.

1. Introduction

Temps et langage sont étroitement liés (Brunet, 1993). La lexicométrie a contribué à mieux comprendre cette relation depuis les travaux d'A. Salem (1988) qui ont défini la notion de série chronologique. Ces recherches initiales se sont ensuite développées dans de nombreux domaines pour souligner l'existence d'un temps lexical. L'analyse de discours politiques (Mayaffre, 2000 ; de Sousa, 2012 ; Ratinaud et Marchand, 2015 ; Diwersy et Luxardo, 2016)

montre une association entre le temps et le contenu du langage utilisé à des échelles historiques. Des études se sont aussi attachées à dégager des tendances de l'évolution du langage dans le temps (Wang et al., 2014 ; Dugué et al., 2016 ; Joselin et al., 2014). Au-delà des méthodes de traitement utilisées, la question de la visualisation a particulièrement mobilisé les chercheurs comme, par exemple, les diagrammes alluviaux (Rosvall et Bergstrom, 2010). Cette rapide synthèse montre un usage important de la lexicométrie dans des domaines différents pour comprendre des évolutions du langage dans le temps. Or, l'association entre le temps et le discours peut aussi concerner le temps dans le discours (Klein, 2008). Cette problématique cherchera alors à repérer les caractéristiques temporelles d'un corpus à travers ses spécificités comme le temps des verbes, certaines formes grammaticales, les adverbes de temps ou les principes de discours. Cet article propose d'aborder cette problématique du temps dans le discours pour compléter les analyses des caractéristiques d'un corpus massif de textes.

La littérature lexicométrique sur le temps dans le discours est moins prolifique que celle sur le discours dans le temps. L'analyse des temps verbaux, particulièrement en français, est complexe et difficilement automatisable. De même, les ensembles de formes associables à des questionnements temporels sont rares et peu établis scientifiquement. Quelques auteurs ont cependant proposé des méthodes et outils d'analyse. E. Brunet (1993) met en évidence une évolution des temps verbaux sur une période longue avec un usage simplifié des formes possibles. D. Longrée et al. (2004) argumentent que la répartition des temps verbaux caractérise des auteurs de textes en latin en proposant des méthodes adaptées à ce type de forme grammaticale. O. Kraif et J. Sorba (2018) puis S. Diwersy et al. (2021) étudient des données massives de romans contemporains pour dégager des styles par genre en utilisant le temps comme un des critères d'analyse. Ces recherches soulignent les enjeux d'une exploration du temps dans un discours. Elles permettent, notamment, d'analyser certaines caractéristiques d'un corpus. Dans cet article, nous proposerons un protocole d'observation des caractéristiques temporelles d'un discours. Il sera expérimenté sur un corpus de données massives d'avis de salariés sur leur employeur. Les résultats seront détaillés puis discutés.

2. Protocole d'observation des caractéristiques temporelles d'un discours

Pour contribuer à cette problématique qui relève de la littérature lexicométrique sur l'étude du temps, nous expérimentons un protocole d'observation du temps dans un discours appliqué à un corpus d'avis de salariés sur leur employeur publiés sur l'internet à partir duquel nous proposons une méthode d'analyse du temps dans le discours.

2.1. Protocole d'analyse du temps dans le discours

Le protocole suit les étapes suivantes (fig. 1) en s'inspirant des travaux de J.M. Leblanc (2015): analyse de l'index hiérarchique à partir du corpus lemmatisé puis non lemmatisé (connecteurs et modalisateurs temporels, champ lexical du temps, temps verbaux) ; analyses contrastives (TGEN sur C.D.H. pour les formes temporelles), relations entre fréquences des formes temporelles et analyse contrastive ; étude des spécificités selon les variables du corpus ; synthèse des caractéristiques temporelles du corpus. Cette expérimentation utilise plusieurs logiciels : Tropes (Ghiglione et al., 1998), Iramuteq (Ratinaud, 2009) et TreeTagger (Schmid, 1995) et les méthodes statistiques suivantes : fréquences des formes, types généralisés (Lamalle et Salem, 2002), classification descendante hiérarchique (Reinert, 1983). Chaque étape sera détaillée dans la section suivante présentant les résultats.

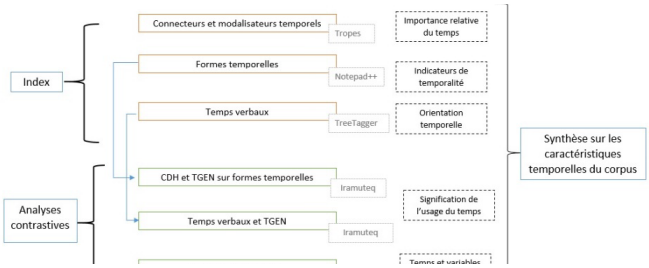


Figure 1. Synthèse du protocole d'observation des caractéristiques temporelles d'un discours

2.2. Présentation du corpus

Le corpus étudié contient des avis publics et anonymes de salariés sur leur employeur collectés sur une plateforme internet. Chaque avis est associé à une note de 1 à 5 comme sur Tripadvisor. Exemple d'avis : «La politique d'évolution interne est très intéressante, c'est un lieu où l'on est en perpétuel mouvement». Nous avons retenu les entreprises pour lesquelles plus de 50 avis avaient été publiés. Le corpus final est de 118 602 avis publiés pour 429 entreprises après nettoyage (avis en langue étrangère, doublons, etc.) contenant 4 451 793 occurrences pour 48024 formes et 38 formes par avis. Ce processus de nettoyage et d'adaptation du corpus à notre problématique s'est étalé de février à mai 2021.

3. Résultats de l'expérimentation du protocole

Cette section présentera les principaux résultats du protocole suivi et appliqué au corpus utilisé pour cette expérimentation.

3.1 Analyse de l'index hiérarchique temporel

Cette première phase de la méthode consiste à analyser les grandes tendances temporelles présentes dans le corpus. Les approches abordent successivement (fig. 1) : le temps en référence à d'autres catégories d'analyse (le corpus est-il plus ou moins temporel ?), le vocabulaire du temps (quel type de vocabulaire temporel est présent dans le corpus ?), les temps verbaux (quel type de temps domine le discours (passé/présent/futur) ?).

3.1.1. La catégorie temporelle dans le discours

Cette phase a pour finalité de fournir des caractéristiques temporelles du discours. Pour cela, un comptage des formes préalablement classées par catégories est réalisée à l'aide du logiciel Tropes. Les discours de ce corpus utilisent peu les connecteurs temporels au sens de ce logiciel (connecteurs temporels : 2,99% des connecteurs). L'orientation argumentative explique la domination des connecteurs d'addition, d'opposition ou de cause. Les formes temporelles servent plus à nuancer, modaliser les avis (modalisateurs temporels : 10,96% des modalisateurs) loin cependant des modalisateurs d'intensité qui sont majoritaires. Ce corpus est ainsi faiblement temporel selon cette première approche globale.

3.1.2. Les formes temporelles du discours

Bien que peu temporels, ces avis de salariés utilisent du vocabulaire associé au temps. Les listes de formes temporelles sont difficiles à établir, sans doute car la catégorisation de mots aux sens multiples est toujours sujette à caution. Nous avons ainsi établi deux listes avec des sources différentes. La première liste comporte 52 connecteurs temporels indiquant la simultanéité, la fréquence, la postériorité et l'antériorité[1]. La deuxième liste est issue des travaux de O. Kraif et J. Sorba (2018)[2] et comprend 42 formes associées au temps que nous avons catégorisées en période, moment, fréquence, antériorité, événement, situation, découpage et saison. Ces listes sont complémentaires car elles abordent des

[1] à peine, à présent, après, au fur et à mesure, au moment, aujourd'hui, auparavant, aussitôt, autrefois, avant, bientôt, cependant, déjà, demain, depuis, derechef, dès que, désormais, en attendant, enfin, ensuite, hier, jadis, jamais, jusqu'à, le temps que, longtemps, lors, lorsque, maintenant, naguère, parfois, pendant, postérieurement, puis, quand, quelquefois, sitôt, soudain, souvent, sur ces entrefaites, sur-le-champ, tandis, tant, tantôt, tard, tôt, toujours, tout à coup, tout à l'heure, tout de suite, ultérieurement. Liste établie à partir du dictionnaire Le Robert et de la publication de J.-M. Kalmbach (2020).

2 an, année, après-midi, aujourd'hui, automne, autrefois, brièvement, décennie, demain, fois, heure, hier, hiver, immédiatement, instant, jadis, jamais, jour, journée, lendemain, longtemps, maintenant, matin, millénaire, minute, mois, moment, nuit, parfois, printemps, seconde, semaine, siècle, soir, soirée, souvent, subitement, tard, temps, tôt, toujours, tout à coup, tout de suite. La forme « été » n'a pas été retenue en raison des confusions entre le verbe être et la saison.

thématiques temporelles différentes, certaines formes sont cependant communes aux deux listes. Les occurrences de chaque forme ont été comptabilisées, les formes retenues représentant environ 90% du total de chaque liste.
La liste 1 met en évidence principalement des connecteurs de simultanéité (42,22%) comme 'quand' ou 'pendant' et des connecteurs de fréquence (33,45%) comme 'toujours' ou 'jamais'. Ces types de formes sont aussi très présents dans la liste 2 (27,92%). Ce deuxième ensemble met surtout en avant les formes de découpage du temps (39,43%) avec 'journée' principalement ou encore 'mois' et 'semaine'.
Ce relevé montre ainsi un usage des formes temporelles favorisant un découpage en journée de travail, en description d'actions simultanées et en soulignant une fréquence très variable, de la rareté à la profusion.

3.1.3. Les temps verbaux
Le temps des verbes est un marqueur temporel important dans un discours mais difficile à analyser sur des données massives. Les confusions sont possibles, par exemple : ' j'ai adoré allez travailler' (erreur comprise) sera difficilement interprété par un algorithme. Afin d'obtenir des résultats assez satisfaisants et automatiques, nous avons utilisé TreeTagger[3]. Pour éviter des interprétations peu précises, nous avons retenu seulement les analyses ayant une probabilité supérieure à 80% ce qui a permis d'interpréter 82% des verbes de ce corpus. Le résultat montre une orientation très majoritaire vers le présent avec quelques rares discours au passé. Malgré la précision des mesures et les précautions prises, ces données doivent être interprétées avec prudence. Des erreurs d'interprétation demeurent ('missions' est analysé comme un subjonctif imparfait et pas un nom) mais la tendance est cependant suffisamment claire pour permettre des analyses. Le croisement de ces orientations générales avec des sous-catégories du discours permettra de faire une analyse contrastive, mieux située dans le contexte du corpus.

3.2. Analyses contrastives du discours
L'analyse contrastive a croisé les deux listes de formes temporelles (2.1.2) avec une classification descendante hiérarchique grâce aux types généralisés et à des variables associées aux discours (analyse de spécificités).

[1] Tags utilisés : VER:pres (présent), VER:ppre (participe présent), VER:subp (subjonctif présent), VER:futu (futur), VER:impf (imparfait), VER:simp (passé simple), VER:subi (subjonctif imparfait), VER:cond (conditionnel), VER:impe (impératif)
[4] Dans ce cas, nous avons obtenu 11 classes regroupant 92,45% des segments du corpus.

3.2.1. Classification descendante hiérarchique et type généralisé

Les deux listes de formes temporelles, limitées aux formes les plus fréquentes ont été croisées avec une classification descendante hiérarchique[4] (Reinert, 1983) par un type généralisé (Lamalle et Salem, 2002). Les résultats permettent de connaitre le degré d'association des formes temporelles avec chaque classe. Cette analyse distingue plusieurs types de classe au filtre de cette analyse temporelle. Certaines comme la classe 6 utilisent un vocabulaire non temporel. La classe 6 correspond à des avis de stagiaires très courts qui énoncent rapidement leur satisfaction. Au contraire, d'autres classes comme les 7, 8 et 9 contiennent des formes temporelles. Globalement, les résultats des deux TGEN présentent des résultats similaires en mettant en évidence ces trois dernières classes. Il existe cependant des nuances. Par exemple, la classe 8 qui critique les salaires utilise des formes temporelles de découpage, le mois en particulier, ou la classe 9 avec la journée. Les raisons de ces différences peuvent être liées à la longueur des avis regroupés dans les classes. Un discours long aura plus de chance de contenir des formes temporelles ce qui pourrait correspondre aux classes 7, 8 et 9. La classe 4, très longue aussi, est cependant un contre-exemple. L'hypothèse, pour ce corpus, est que les classes contenant des discours critiques et/ou décrivant une action sont des récits, des narrations intégrant des logiques temporelles.

3.2.2. Temps verbaux et TGEN

Le croisement des temps verbaux avec les différentes classes souligne les caractéristiques de quelques regroupements. Pour montrer des différences, nous avons centré les analyses sur les deux temps principaux, présent et passé. L'usage du temps passé correspond à des classes présentant le contenu d'emploi, la description d'une journée type pour les classes 2, 9 et 11. L'usage du passé pour la classe 6 s'explique plutôt par le type de narrateur, des stagiaires. Les autres classes sont dans la moyenne générale. On peut cependant noter que l'usage de l'impératif est beaucoup plus fréquent dans les classes 4, 7 et 8 qui contiennent des discours critiques avec, notamment, l'expression 'Fuyez !'.

3.2.3. Spécificités du discours et formes temporelles

Cette dernière étape consiste à croiser les listes de formes temporelles (listes 1 et 2) avec les variables associées aux discours étudiés. Dans ce cas, nous avons choisi la variable des notes et deux formes associées à la fréquence, la seule catégorie commune aux deux listes. Ces deux formes ont des fréquences opposées qui sont associées à des notes également différentes. La forme 'toujours' est associée à des notes élevées comme la note 5 alors que 'jamais' est plus associée à la note 2. Les avis associés à la note 5 tendent à généraliser une interprétation positive de leur employeur alors que les notes faibles soulignent des absences. Ce croisement aurait pu aussi se faire avec des variables

temporelles ce qui permettrait une analyse du temps dans le discours… dans le temps.

4. Discussions et conclusion

Le protocole expérimenté dans cet article a permis de mieux comprendre une partie du sens des discours. L'utilisation massive du présent montre que les auteurs sont, *a priori*, des employés actuels. L'approche montre aussi que certains discours intègrent peu les formes temporelles. Enfin, il existe un lien entre le contenu du discours et ses caractéristiques temporelles comme l'ont montré O. Kraif et J. Sorba (2018) puis S. Diwersy et al. (2021) pour les romans. La critique des employés pour ce corpus est structurée temporellement et propose une forme plus narrative que d'autres avis enthousiastes associés aux logiques de marque employeur. L'apport principal est cependant de proposer un travail d'élaboration d'une méthode d'observation du temps dans le discours. A ce titre, il a de nombreuses limites. D'abord, le corpus choisi est faiblement temporalisé ce qui n'a pas facilité le travail sur ce questionnement. D'autres corpus plus variés pourront répliquer, compléter et affiner cette première approche. Ensuite, le choix des mots temporels est une étape importante mais délicate et orientant les explications suivantes. De plus, la détermination du temps des verbes, comme souligné précédemment, est améliorable avec de nouveaux outils comme des algorithmes intégrant de l'apprentissage supervisé. De même, nous n'avons pas étudié les liens entre temps verbaux, comme variable explicative, et classification par des tests du Chi². Ces trois limites importantes (choix des formes, temps verbaux, variable) ouvrent de nombreuses pistes de recherche futures.

Bibliographie

Brunet E. (1993). Quand le temps change avec le temps. *Texto ! Textes et Cultures*, XXI(1), 1-27.

Diwersy S., Gonon L., Goossens V., Kraif O., Novakova I., Sorba J. et Vidotto I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, 22, 1-23.

Diwersy S. et Luxardo G. (2016). Mettre en évidence le temps lexical dans un corpus de grandes dimensions : L'exemple des débats du Parlement européen. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 1-12.

Dugué N., Lamirel J.C. et Cuxac P. (2016). Visualisation pour la détection d'évolutions dans des corpus de publications scientifiques Indexation, classification et analyse diachronique pour la visualisation. *Les Cahiers du Numérique*, *4*, 157-183.

Ghiglione R., Landré A. et Molette P. (1998). *L'Analyse automatique des contenus*. Dunod.

Joselin L., Eliot E., Jeanne P., Lepastourel N., Gasquet C. & Amalric M. (2014). Dynamiques temporelles de la pandémie de grippe A/H1N1 dans la presse écrite francophone. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 1-12.

Kalmbach J.M. (2020). *Guide de grammaire française pour étudiants finnophones*. Institut des Sciences du langage et de la communication Université de Jyväskylä Finlande.
Klein W. (2008). Time in Language, Language in Time. *Language Learning*, *58*(Supplément 1), 1-12.
Kraif O. et Sorba J. (2018). Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale). *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 392-399.
Lamalle C. and Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Journées internationales d'analyse statistique des données textuelles (JADT*), 403-411.
Leblanc J.M. (2015). Proposition de protocole pour l'analyse des données textuelles : Pour une démarche expérimentale en lexicométrie. *Nouvelles perspectives en sciences sociales*, *11*(1), 25-63.
Longrée D., Luong X. et Mellet S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle : Analyses d'un corpus lemmatisé. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 743-752.
Mayaffre D. (2000). Temps lexical ou temps politique ? *Lexicometrica*, 1-10.
Ratinaud P. (2009). *Iramuteq* (Version 0.7 Alpha 2) [Linux, Windows, MacOS].
Ratinaud P. et Marchand P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, *108*(2), 57-77.
Reinert M. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, *8*(2), 187-198.
Rosvall M. et Bergstrom C. (2010). Mapping Change in Large Networks. *PLoS ONE*, *5*(1), 1-7.
Salem A. (1988). Approches du temps lexical Statistique textuelle et séries chronologiques. *Mots*, *17*, 105-143.
Schmid H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, 1-9.
Sousa (de) S. (2012). À l'épreuve des temps… Temps lexical et temps politique dans le discours de Fidel Castro (1959-2008). *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 337-349.
Wang X., Cheng Q. et Wei L. (2014). Analyzing evolution of research topics with NEViewer : A new method based on dynamic co-word networks. *Scientometrics*, *101*, 1253–1271.

Saisir par la textométrie la notion de radicalisation dans des corpus institutionnels hétérogènes

Manon Pengam[1], Agata Jackiewicz[1], Pascal Marchand[2]

[1]Université Paul-Valéry Montpellier 3 – manon.pengam@univ-montp3.fr agata.jackiewicz@univ-montp3.fr
[2]Université Toulouse 3 Paul Sabatier – pascal.marchand@iut-tlse3.fr

Abstract
This article offers an exploratory and contrastive study of two heterogeneous institutional corpus on radicalization: public speeches by the executive (2013-2018) and reports from parliamentary inquiry commissions (2013-2017). Using the IRaMuTeQ software, we apply to each corpus a descending hierarchical classification (CHD) according to the Reinert method (1983, 1990). The projection of the lemmas *radical* and *radicalization* on the classes reveals which is the "lexical world" which primarily invests the theme of jihadist radicalization in the two genres of discourse studied. The interpretive path is supplemented by a second CHD on this majority class and a chronological variable which makes it possible to contextualize the results observed.

Keywords: institutional speeches, radicalization, textometry, IRaMuTeQ

Résumé
Cet article propose une étude exploratoire et contrastive de deux corpus institutionnels hétérogènes sur la radicalisation : des déclarations publiques de l'exécutif (2013-2018) et des rapports issus de commissions d'enquêtes parlementaires (2013-2017). À l'aide du logiciel IRaMuTeQ, nous appliquons sur chaque corpus une classification hiérarchique descendante (CHD) selon la méthode Reinert (1983, 1990). La projection des lemmes *radical* et *radicalisation* sur les classes révèle quel est le « monde lexical » qui investit prioritairement le thème de la radicalisation djihadiste dans les deux genres de discours étudiés. Le parcours interprétatif est complété d'une deuxième CHD sur cette classe majoritaire et d'une variable chronologique qui permet de contextualiser les résultats observés.

Mots-clés: discours institutionnels, radicalisation, textométrie, IRaMuTeQ

1. Introduction

Dès le mitant des années 2010, la problématique de la radicalisation djihadiste a

généré une importante production de discours institutionnels français et une série de plans d'actions politiques destinés à éviter la commission d'actes terroristes sur le territoire national.

Notre travail, situé au confluent de l'analyse de discours et de la statistique textuelle, propose de comparer deux types de discours institutionnels français sur la radicalisation :

- des déclarations publiques de l'exécutif (2013-2018), issus de Pengam (2021), désormais corpus DECL (2 millions d'occurrences, 680 déclarations)
- des rapports provenant des commissions d'enquêtes parlementaires (2013-2017), issus de Marchand (2017), désormais corpus RAPP (1 million d'occurrences, 5 commissions d'enquête)

En première lecture, ces deux corpus possèdent des orientations argumentatives différentes et des caractéristiques énonciatives propres. Les rapports d'enquêtes parlementaires prescrivent des recommandations (menant parfois à des propositions de loi), qui émanent du législateur mais aussi d'acteurs de la société civile (syndicalistes de l'administration pénitentiaire, témoins, avocats…) auditionnés en raison de leur expertise à l'égard de la problématique considérée. Les déclarations publiques de l'exécutif (ministres, Premier Ministre, Président de la République) imposent, quant à elles, une vision homogène et consensuelle des problèmes publics, nivelant la diversité des points de vue exprimés.

Dans un contexte d'urgence sociale et politique liée aux attentats terroristes, dans quelle mesure – et si oui comment – les recommandations émises par les commissions d'enquête sont-elles reprises par l'action publique ? Plus généralement, comment s'est construite et a évolué la représentation publique de la radicalisation djihadiste en France ?

Pour tenter de répondre à ces questions qui s'inscrivent dans un programme de travail échelonné en plusieurs étapes (Pengam et *al.* à paraître), nous réalisons dans un premier temps une étude exploratoire contrastive de ces deux ensembles discursifs, par classification hiérarchique descendante (Reinert 1983, 1990) à l'aide du logiciel IRaMuTeQ, et dont nous livrons les résultats dans l'article.

2. Premières explorations des corpus par classification hiérarchique descendante

2.1. Analyse du corpus DECL

Une première analyse par classification hiérarchique descendante sur le corpus DECL a permis d'en identifier les principales classes de discours, [Figure 1]. Un premier bloc met en exergue la thématique des valeurs républicaines (classe 1 : *république*, *valeur*, *rassembler*…) et des relations internationales dans un contexte marqué par l'émergence de l'organisation État Islamique (classe 5 : *Syrie, France, Europe, Daech…*). Face à cette menace terroriste, est convoquée

une double réponse : à la fois législative (classe 4 : *loi*, *procédure*, *contrôle*, *mesure*…), et de renseignement et de contrôle sur les populations (classe 3 : *service*, *renseignement*, *sécurité*…). Deux classes de discours semblent plus isolées : la première porte sur le secteur de l'économie et de l'innovation (classe 2 : *entreprise*, *investissement*, *développement*…), la seconde, plus surprenante, est constituée des patronymes des représentants politiques et des journalistes (classe 6 : *jean*, *valls*, *bourdin*…). Cette surreprésentation s'explique en réalité par la présence importante du genre de l'interview politique dans le corpus DECL.

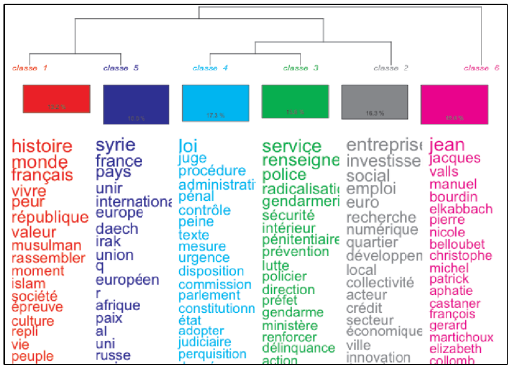


Figure 1. Classes de discours du corpus DECL

La projection des lemmes *radical* et *radicalisation* sur les six classes issues du dendogramme 1 permet de mieux saisir quelle classe investit prioritairement le thème de la radicalisation djihadiste. En l'occurrence, comme l'illustre la figure 2, il s'agit de la classe 3, qui touche au domaine de la sécurité intérieure.

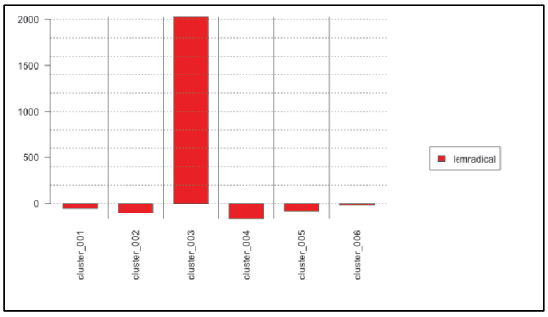


Figure 2. Projection des lemmes radical et radicalisation sur les classes du corpus DECL

Nous avons appliqué une nouvelle classification, spécifique à la classe 3 cette fois, afin d'approfondir ses univers de discours. Deux grands groupes thématiques se distinguent. Le premier se rapporte à la gestion sécuritaire du terrorisme, à la fois dans les différents services garants du maintien de l'ordre : (classe 5 : *pénitentiaire*, *renseignement*, *dgsi*…), et dans les opérations concrètes des forces de l'ordre, en particulier militaires (classe 2 : *force*, *gendarme*, *militaire*…). Le deuxième groupe concerne les moyens mis en œuvre pour contrer la propagande djihadiste sur Internet (classe 4 : *internet*, *propagande*, *numérique*…), et le démantèlement de filières criminelles (classe 3 : *lutte*, *trafic*, *délinquance*, *criminalité*, *démanteler*…). La classe 6 place la focale sur les individus suspectés de radicalisation, dont il s'agit de suivre la trajectoire (*suivre*, *radicalisés*, *signaler*…).

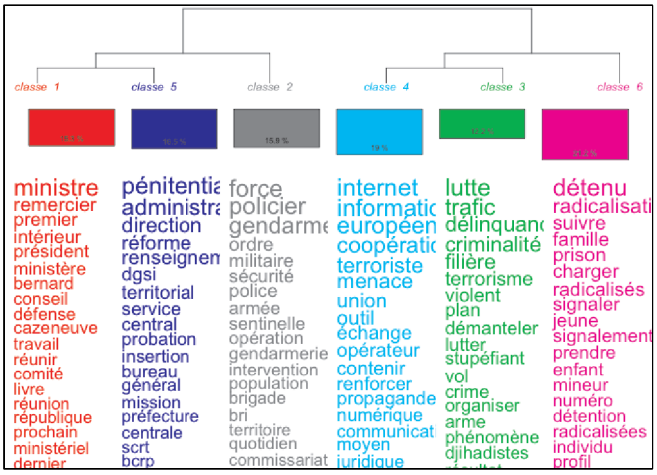


Figure 3. Dendogramme de la classe 3 « sécurité intérieure » (corpus DECL)

Les thématiques sont temporellement marquées et révèlent des changements dans les représentations des causes jugées responsables par le politique de la radicalisation. En 2014, la priorité est placée sur le démantèlement des filières criminelles de tout type (classe 3 : délinquance, criminalité, stupéfiant, vol…), qui ont un lien potentiel avec l'entrée dans la militance djihadiste. En 2015 c'est le thème du numérique qui apparaît comme un vecteur de l'engagement terroriste (classe 4). En 2013 et 2017, le thème de la gestion sécuritaire du terrorisme est significativement présent. Nous faisons l'hypothèse qu'il s'agit pour l'exécutif nouvellement au pouvoir (2013 : gouvernement Ayrault, 2017 : gouvernement Philippe) d'afficher ses objectifs en matière de lutte contre le terrorisme. En 2016, l'accent est mis sur l'individu radicalisé.

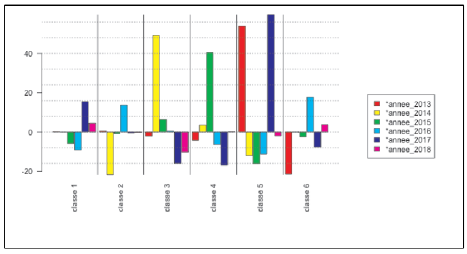


Figure 4. Évolution temporelle des thématiques de la classe 3 « sécurité intérieure » (corpus DECL)

Analyse du corpus RAPP

De la même façon, nous avons appliqué une classification hiérarchique descendante sur le corpus RAPP. Sans surprise, celui-ci présente une plus grande homogénéité dans les thématiques abordées. Les rapports ici analysés sont en effet issus de commissions d'enquêtes *sur* la radicalisation, là où le corpus DECL rassemble des déclarations publiques ne portant pas uniquement sur cette thématique. Les classes de discours du corpus RAPP portent sur les récits d'intervention de la sécurité publique et des secours (classe 6 : *BRI*, *raid*, *intervention*, *secours*…), la radicalisation djihadiste (classe 2 : *musulman*, *religieux*, *radicalisation*…), la réponse législative que ce phénomène appelle (classe 1 : *article*, *code*, *peine*…), les domaines du renseignement (classe 3 : *renseignement*, *service*, *dgsi*…) et de l'univers pénitentiaire (classe 4 : *pénitentiaire*, *cellule*, *établissement*…) amenés à garantir le maintien de l'ordre public, et le lexique spécialisé propre à la gestion de la commission elle-même (classe 5 : *commission*, *enquête*, *audition*…).

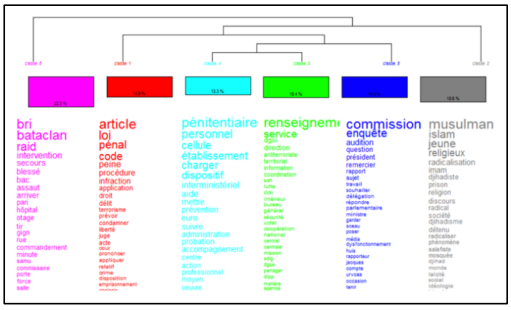


Figure 5. Classes de discours du corpus RAPP

La projection des flexions correspondant au lemme *radical* dévoile un investissement plus important de l'univers de discours radicalisation dans la classe 2, relative à la radicalisation djihadiste.

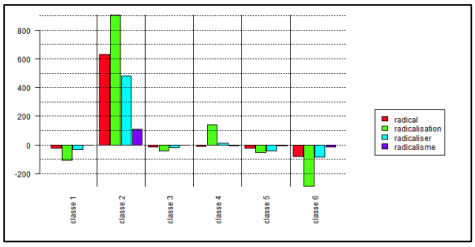


Figure 6. Projection du lemme radical dans la classe 2 « radicalisation djihadiste » (corpus RAPP)

Une deuxième CHD effectuée sur la classe 2 fait apparaître deux grands groupes thématiques. Le premier se rapporte à la question religieuse et ses rapports avec la République (classe 1 : *républicain*, *chrétien*, *voile*, *laïcité*…), à l'extrémisme islamiste (classe 2 : *salafisme*, *quiétiste*…), et plus généralement à la gestion du culte musulman (classe 5 : *culte*, *aumônier*, *enseignement*…). Le second groupe concerne les individus dont les trajectoires impliquent un séjour en prison (classe 7 : *détenu*, *prison*, *carcéral*…), Internet (classe 6 : *réseau*, *internet*, *diffuser*…), une adolescence marquée par une quête djihadiste (classe 3 : *jeune*, *fille*, *psychologique*…), et enfin le passage à l'acte terroriste (classe 4 : *attentat*, *Merah*, *attaquer*…).

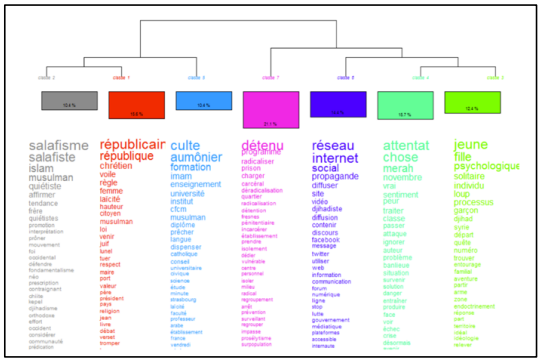


Figure 7. Dendogramme de la classe 2 « radicalisation djihadiste » (corpus RAPP)

Là encore, les thématiques évoluent chronologiquement. La commission du Sénat de 2015 place la focale sur Internet (classe 6). La même année, la commission de l'Assemblée investit prioritairement les valeurs de la République (classe 1). La question du passage à l'acte terroriste (classe 4) semble au cœur de la commission de l'Assemblée de 2016.

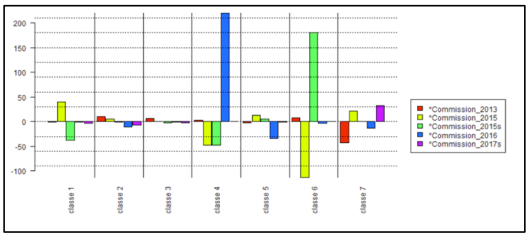


Figure 8. Évolution temporelle des thématiques de la classe 2 « radicalisation djihadiste » (corpus RAPP)

Bilan

Les CHD réalisées éclairent les représentations que le politique et le corps parlementaires accolent à la radicalisation djihadiste, et l'évolution chronologique des thématiques entre 2013 et 2018. On constate en particulier une focalisation, dans les deux corpus, sur le numérique comme vecteur de l'engagement djihadiste en 2015, et une approche individuelle du passage à l'acte en 2016. Si les CHD réalisées ne mettent pas en évidence les procédés argumentatifs et énonciatifs qui portent ces thèmes, elles dessinent de nouvelles pistes d'analyse qui, une fois investiguées, permettraient de caractériser et de différencier plus finement les deux sources de discours étudiés. Par exemple : la présence en 2016, dans le corpus de commissions d'enquêtes parlementaires, du thème des valeurs républicaines, nous interroge sur le statut de leur(s) énonciateur(s) : décideurs politiques membres de la majorité présidentielle ? de l'opposition ? représentants associatifs de la société civile ?

3. Perspectives

En seconde perspective de comparaison des corpus DECL et RAPP, nous analyserons (travail en cours) les deux sous-corpus constitués par extraction de segments de texte à partir du lemme pivot *radical*. Nous chercherons à mesurer leur distance/proximité textuelle à partir d'une analyse factorielle des correspondances sur le tableau lexical complet (Lebart et Salem 1994) et une mesure des distances textuelles avec l'indice de Labbé (Labbé et Labbé 2003). Plus largement, nous souhaitons par ce travail qui s'inscrit dans une démarche générale d'analyse des notions socio-politiques complexes et sensibles (Jackiewicz et Pengam 2020 ; Ratinaud et Marchand 2012), fournir une contribution à l'analyse des problèmes publics au sens de Neveu (2017), tels qu'ils sont représentés par la parole institutionnelle (Née, Oger et Sitri 2017 ; Krieg-Planque 2012 ; Oger et Ollivier-Yaniv 2003).

Bibliographie

Jackiewicz A. et Pengam M. (2020). Un modèle pour l'étude des nominations émergentes. Notion de repérage pour saisir les modalités d'ajustement sémantique et discursif. *SHS Web of Conferences, 78, Congrès Mondial de Linguistique Française 2020*. https://doi.org/10.1051/shsconf/20207812004

Krieg-Planque A. (2012). *Analyser les discours institutionnels*. Armand Colin.

Labbé C. et Labbé D. (2003). La distance intertextuelle. *Corpus*, (2) : 95-117.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris, France. Dunod.

Marchand P. (2017). « La fabrique parlementaire du discours sur la "radicalisation" : politiques, acteurs, experts », dans Baygert Nicolas, Durin Esther, Maas Élise et Nicolas Loïc (dirs.), *Communiquer (sur) la radicalité*, Les Cahiers PROTAGORAS, 4, 2017, 30-46.

Marty E., Marchand P. et Ratinaud, P. (2013). Les médias et l'opinion : éléments théoriques et méthodologiques pour une analyse du débat sur l'identité nationale. *Bulletin de Méthodologie Sociologique*, 117(1) : 46–60.

Née É., Oger C. et Sitri, F. (2017). Le rapport : Opérativité d'un genre hétérogène. *Mots. Les langages du politique*, *114*, 9-24.

Neveu E. (2017). L'analyse des problèmes publics. *Idées économiques et sociales, 190(4)*, 6-19.

Oger C. et Ollivier-Yaniv C. (2003). Analyse du discours institutionnel et sociologie compréhensive : vers une anthropologie des discours institutionnels. *Mots. Les langages du politique*, *71*, 125-145. https://doi.org/10.4000/mots.8423

Pengam M. (2021). *Pour une modélisation linguistique de la radicalisation. Étude de discours institutionnels et de discours du travail social*. Thèse de doctorat. Université Paul-Valéry Montpellier 3.

Ratinaud P. et Marchand, P. (2012). Recherche improbable d'une homogène diversité : le débat sur l'identitié nationale. *Langages, 187,* 93-107.

Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les Cahiers de l'analyse des données, 3, 187-198.

Reinert M. (1990). ALCESTE: Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval. Bulletin de Méthodologie Sociologique, 26, 24-54.

Réalité virtuelle: quels apports pour la visualisation en textométrie/lexicométrie ?

Marie Pérès, Jean-Marc Leblanc

Université Paris Est Créteil – marie.leblanc@u-pec.fr jean-marc.leblanc@u-pec.fr

Abstract
This paper focuses on how virtual reality could make significant contributions to the visualization of results, but also to the exploration of data in TDA as a facilitator for the extraction of new knowledge. These technologies, particularly adapted to collaborative work, also facilitate dialogue for study or diagnostic purposes. The virtual reality that we are calling upon here aims to allow the scientist to adopt a new point of view, new approaches to data exploration (2D/3D/4D) in order to improve the exploration and understanding of large volumes of data and thus open the field of perceptual data mining.

Keywords: Textual Data Analysis, Virtual Reality, Augmented reality, Perceptual mining

Résumé
Cet article s'attache à déterminer en quoi la réalité virtuelle pourrait permettre des apports significatifs dans les visualisations de résultats mais aussi dans l'exploration des données en ADT en tant que facilitateur d'extraction de nouvelles connaissances. Ces technologies particulièrement adaptées au travail collaboratif facilitent également le dialogue à des fins d'études ou de diagnostic. La réalité virtuelle que nous convoquons ici a pour but de permettre au scientifique d'adopter un nouveau point de vue, de nouvelles approches d'exploration des données (2D/3D/4D) afin d'améliorer l'exploration et la compréhension de grands volumes de données et ouvrant ainsi le champ de la *fouille perceptuelle des données*.

Mots clés: Analyse de Données Textuelles, Réalité Virtuelle, Réalité Augmentée, Perceptual mining

1. Introduction

Si la visualisation tridimensionnelle permet d'interpréter certaines configurations des Analyses Factorielles des Correspondantes (AFC) produites en ADT (analyse statistique des données textuelles), d'autres moyens technologiques peuvent constituer des apports significatifs dans les visualisations de résultats mais aussi dans l'exploration des données.
Cet article s'intéressera aux types de dispositifs communicationnels pouvant être mis en place pour l'exploration, l'interaction et la co-habitation multi-sensorielle

en réalité virtuelle pour l'ADT.
Si une réflexion a déjà été menée dans nombre de domaines scientifiques soulevant de nouvelles questions aux concepteurs d'environnements virtuels, l'application de la réalité virtuelle à la lexicométrie ne semble pas en faire partie. Depuis 20 ans, de nombreux exemples en architecture, archéologie, biologie moléculaire ou en physique ont été développés ou sont en cours de développement. Cet article s'attachera donc à poser les bases des questionnements et pistes qui s'ouvrent à nous dans la poursuite de cette exploration dans le champ des ADT.

2. Data mining et Visual mining: les potentialités de la RV

L'exploitation des résultats de l'analyse statistique de données textuelles rejoint par nombre d'aspects le « data mining » ou « fouille de données » celle-ci consistant en l'extraction de données organisées et de motifs pertinents depuis des données brutes, en général de grande taille (Big data) afin de pouvoir les exploiter pour se saisir d'un ensemble impossible à percevoir autrement. De même l'ADT permet de mieux se saisir de corpus textuels en utilisant des outils statistiques voire pour certains d'éléments relevant du deep learning. L'interdisciplinarité est au cœur du développement de ces outils, le scientifique étant acteur dans la boucle de simulation, partenaire dans le développement et utilisateur-évaluateur en aval.
De la même manière, le « visual mining » peut être abordé selon deux méthodes (Azzag et al. 2008). La première, très automatisée, implique l'intervention d'un expert lors du codage initial des informations visuelles sous forme textuelle interprétable par les outils mais aussi en phase finale pour l'interprétation. Cette démarche permet ainsi d'analyser des images en combinant des logiciels d'aide à l'analyse qualitative (CAQDAS – Computer Assisted Qualitative Data Analysis Software) et des outils d'ADT utilisés en tant qu'outils de fouille visuelle de données (Pérès 2016). Le second, consiste en un processus de fouille de données avec représentation desdites données par des signes visuels visant à permettre l'interprétation visuelle directe et facilitant ainsi l'accomplissement de la tâche à réaliser par l'expert tout en permettant l'interaction souvent intuitive entre l'expert et la visualisation (notion de requête graphique) (Wong and Bergeron 1994). Si les deux axes relèvent finalement de procédés très proches dans le cas d'une analyse de corpus d'images, la seconde a l'avantage de ne pas nécessiter un encodage textuel de données visuelles. Dans le cas de corpus textuel la transposition se ferait du texte vers un visuel voire de résultat statistique en visuel permettant ainsi de manière plus directe une « fouille visuelle de données ».
Cette démarche de conception est à rapprocher de celle des outils de type « Visual analytics » que l'on peut traduire par « outils de visualisation analytiques » (Shrinivasan and van Wijk 2008). Ce champ de la visualisation de données se

concentre sur l'utilisation de visualisations (le plus souvent interactives) pour faciliter le raisonnement analytique le plus souvent assisté d'une IA (deep learning).
Depuis quelques années, l'apparition du terme « immersive analytics » (Chandler et al. 2015) reflète bien l'importance que revêt l'immersion dans ce champ avec, par exemple, le développement des travaux du IALab de la Monash University de Melbourne[1] (Yang et al. 2020) (Skarbez et al. 2019)
La distinction ne se situe donc pas tant entre « data mining » et « visual mining » mais sur la forme ou les formes de représentation permettant de se saisir des données à fouiller qu'elles soient textuelles, visuelles (images fixes ou vidéo), audio, multimédias ou flot de données.
La réalité virtuelle que nous convoquons ici a pour but de permettre au scientifique d'adopter un nouveau point de vue, de nouvelles approches d'exploration des données (2D/3D/4D) afin d'améliorer l'exploration et la compréhension de grands volumes de données et ouvrant ainsi le champ de la *fouille perceptuelle des données*.

3. Travail collaboratif: les apports potentiels de la RV en ADT

La collaboration (éventuellement à distance) que permettent les technologies de RV facilite le dialogue à des fins d'études ou de diagnostic. En effet, comme le démontre Narayan (Narayan et al. 2005) dans une étude sur le degré d'immersion et son influence sur la convivialité dans un contexte de travail collaboratif en binôme, l'immersion améliore grandement la communication qui se construit dans ce contexte. En utilisant le rôle de médiation de la réalité virtuelle dans une pratique scientifique de travail collaboratif, les utilisateurs peuvent voir leurs résultats améliorés de façon significative. Cette étude compare des résultats obtenus avec différents moyens d'immersion (CAVE, casque de réalité virtuelle, avec ou sans son immersif, ou détection du mouvement de la tête) et montre que le temps nécessaire à l'achèvement de la même tache se réduit au fur et à mesure de l'augmentation de l'immersion du médium analysé.
L'exemple du Studierstube, environnement collaboratif en réalité augmentée (Szalavári et al. 1998) (Schmalstieg et al. 2002), illustre également les possibilités de l'immersion mais aussi celles de l'interaction indépendante sur les éléments virtuels manipulés. En effet, cette étude souligne que dans une CAVE seul le « pilote » manipule réellement l'interface virtuelle, les autres participants restant spectateurs. De la même manière, dans le cas d'un environnement immersif

[1] https://ialab.it.monash.edu/current-projects/?marvl_project=19

l'indépendance d'action de chacun des participants et l'interactivité qui en découle (toute action d'un des participants modifie le modèle scientifique immédiatement) augmente de manière significative l'efficacité de la collaboration.
Le travail collaboratif d'analyse de corpus est courant en ADT. Cependant, malgré les possibilités de créer des rapports (Alceste par exemple) facilitant l'échange sur les résultats, aucun logiciel ne permet de manipuler à plusieurs et dans un même temps l'ensemble des ressources produites sur un corpus et de comparer facilement les résultats des choix opérés.
Dans cette étude les auteurs soulignent également l'importance de l'annotation dans des espaces de collaboration en RA qui, combinés aux graphiques et autres visualisations, permettent de rendre explicites des notions complexes ou des réflexions des utilisateurs explicitant les connaissances ainsi créées et font trace de leur construction dans l'espace collaboratif. Ces annotations peuvent être placées dans l'espace virtuel et attachées aux différentes représentations. Forts de cette même constatation dans l'utilisation d'outils d'ADT non immersifs (TextObserver, Le trameur) nous ne pouvons imaginer un outil immersif d'ADT sans ce type d'implémentation qui renforce la collaboration.

4. Explorer les données : la RV comme facilitateur d'extraction de nouvelles connaissances

Si dans les data-sciences les interfaces de fouille visuelle sont la plupart du temps composées d'un ordinateur pourvu d'une interface utilisateur « classique » combinant écran, souris et clavier, il est parfaitement établi que l'expérience utilisateur dans des systèmes interactifs influence fortement le degré d'engagement et améliore la productivité (Hassenzahl and Tractinsky 2006). De plus, comme l'évoque Burkhardt (Brukhardt 2003), en capturant le mouvement, proposant des représentations visuelles, utilisant des dispositifs de capture et de représentation sonore voire en permettant d'exploiter des dispositifs de retour proprioceptifs et cutanés la réalité virtuelle propose des interfaces comportementales (au sens de Fusch) « visant à exploiter un comportement humain, naturel et sans acquis préalable » (Fusch 1996).
Une interface de fouille visuelle écran permet de visualiser un certain nombre de données qu'elle passe par des visualisations bidimensionnelles ou tridimensionnelles et l'ajout d'une couche d'interactivité permet déjà d'ajouter une dimension perceptive aux objets étudiés. Ainsi, le Zoom dans les AFC proposé dans (Leblanc and Pérès 2011) permet de se saisir de l'information en ajoutant la notion de niveau d'information… implémenter l'immersion dans la visualisation tridimensionnelle de l'AFC permettrait d'affiner la représentation des données en améliorant leur perception. En effet, l'utilisateur adopterait alors un point de vue interne à la représentation tout en continuant de pouvoir s'en

saisir et de la manipuler.
Comme le souligne Yang (Yang et al. 2020), les représentations de données abstraites n'ont pas d'échelle naturelle ce qui facilite la possibilité de zoom mais aussi celle de les considérer de n'importe quel point de vue afin de faciliter le travail d'analyse. Il met au cœur de sa réflexion la notion de *navigation incarnée* soulignant les possibilités des interfaces de Réalité Virtuelle dans l'abord des visualisation analytiques. Il évoque ainsi la « téléportation » implémentée dans beaucoup d'outils de RV, qui est devenue un standard permettant de franchir de grandes distances alors que d'autres outils ne permettent que de se déplacer physiquement pour se déplacer dans l'espace virtuel et privilégient l'utilisation du zoom/dé-zoom pour se saisir de la représentation. Ces deux modes de « déplacement » ont un impact très important sur la perception de l'objet virtuel. En effet, le premier prend l'échelle humaine comme référence et propose donc à l'utilisateur de parcourir la représentation, renforçant l'impression d'immersion en le plongeant dans la représentation, alors que le second place l'utilisateur en position de démiurge qui peut à sa guise agrandir ou rétrécir l'objet représenté qu'il saisit. Ces deux modes sont intéressants et il nous semble qu'il serait pertinent de les implémenter conjointement puisqu'ils permettent à l'utilisateur de considérer les données d'un point de vue très différent et de s'en saisir autrement.
De plus, explorer et interagir en temps réel avec les données est au cœur de la construction des outils d'ADT. L'utilisation de la réalité virtuelle permettrait d'extraire de nouvelles connaissances ou de nouvelles organisations non visibles ou perceptibles par les moyens actuels de visualisation des logiciels.
Comme le souligne également Blum (Blum et al., 2019) dans la plupart des applications de fouille visuelle, l'utilisateur interagit indirectement avec les données en agissant sur des éléments d'interfaçage comme des menus, des curseurs à ajuster, etc… permettant de modifier les paramètres de la visualisation. Les interactions sémantiques en RV permettent à l'utilisateur d'interagir spatialement avec la métaphore visuelle leur donnant plus de sens que l'interaction indirecte sur des éléments d'interfaçage. Ainsi, dans un espace de travail virtuel, l'utilisateur peut réorganiser les documents pour rapprocher les données en cluster par similarités ou liens de relation permettant de faire émerger des relations implicites. De plus, la limitation d'un écran d'ordinateur qui amène à fermer ou réduire certaines fenêtres pour permettre de se saisir d'autres n'a plus lieu d'être. Augmenter radicalement l'espace de visualisation permet de faire des comparaisons visuelles directement en comparant côte à côte, voire en superposant de manière tridimensionnelle des éléments. Les possibilités de la réalité virtuelle permettent ainsi de maximiser les capacités de reconnaissance de motifs et de reconnaissance de formes du cerveau humain rendant plus efficace encore l'interprétation des données (Whitlock 2020).

Ces possibilités sont également très bien illustrées par un outil de réalité virtuelle créé au milieu des années 2000 et visant à simuler des protéines en mettant en lumière les cooccurrences les construisant (Férey et al., 2004 - p102). Dans cet exemple c'est l'importance de la stéréoscopie mais aussi de l'interaction en RV qui fut mis en avant pour permettre de démêler les chaines de protéines complexes. Nous ne pouvons que rapprocher cet interfaçage de ceux déjà envisagés pour la visualisation des graphes de cooccurrences en ADT qui souffrent souvent des limitations d'une interface bidimensionnelle dans la perception et l'analyse de l'ensemble des informations complexes qu'ils contiennent.

De la même manière Lisle (Lisle et al. 2020) insiste sur la difficulté de l'exercice consistant à créer des relations et faire émerger du sens lorsque l'on traite une grande quantité de données multimédia et développe une méthode visant à permettre de figurer la qualité hypothétique d'un rapprochement d'information et s'appuyant sur les possibilités facilitatrices de la RV. On peut aisément faire un parallèle avec le traitement des résultats d'analyse d'ADT comme avec celui de l'annotation évoquée ci-avant afin d'envisager l'implémentation d'un module.

L'outil d'ADT en Réalité Virtuelle que nous envisageons ici utiliserait donc une combinaison de ces fonctionnalités permettant ainsi de comparer plus efficacement encore les représentations de résultat d'analyse mais aussi de créer des rapprochements de visualisations structurés en hypothèses annotées s'appuyant fortement sur la représentation par métaphore visuelle dont les possibilités ne font qu'être effleurées en l'état actuel des recherches dans le domaine.

Bien que les exemples évoqués puissent donner l'impression de ne porter que sur les résultats d'analyse c'est bien le développement d'un logiciel d'ADT en RV que nous envisageons[2]. L'interaction multimodale en entrée logiciel doit également être interrogée. Faire évoluer les outils d'ADT en réalité virtuelle nécessite une exploration des modalités ergonomiques des logiciels (commandes vocales, gestuelles, objets d'interactions, etc…). En effet l'interface en Réalité Virtuelle n'est pas seulement dédiée à visualiser des résultats, les rapprocher ou les annoter mais aussi à utiliser le logiciel d'ADT en RV. L'interfaçage permettant de relancer un calcul après avoir modifié un corpus ou un paramétrage, doit lui aussi être pensé selon ces modalités.

5. Conclusion

L'interaction multimodale sur des visualisations de données complexes relève d'un champ que l'on peut qualifier de « fouille perceptuelle de données » et

[2] La maquette du dispositif proposé est disponible à l'adresse suivante : http://info-text.fr/JADT2022/

nécessite une réflexion profonde sur la mise en forme de ces données pour permettre leur étude. La complexité des données représentées doit offrir un rendu et une interaction multi-sensoriels en combinant les trois canaux sensori-moteurs que sont le visuel, l'audio et l'haptique (réflexion sur les modalités et les types d'information) pour permettre leur meilleure perception et l'interprétation la plus pertinente. La condition est simplement que les données représentées se prêtent à une mise en forme adaptée. Travailler en situation immersive semble effectivement la méthode la plus efficace mais on ne négligera pas les possibilités offertes par les outils de réalité augmentée qui présentent également des atouts indéniables dans ces nouvelles modalités d'ADT.

Bibliographie

Azzag H., Da Costa D., Guinot C. and Venturini G., (2008). *Revue des Nouvelles Technologies de l'Information,* Editions RNTI, vol.(A-2), pp.1-14

Blum S., Cetin G. and Stuerzlinger W. (2019). Immersive Analytics Sensemaking on Different Platforms, Computer Science Research Notes, *Proc. of WSCG 2019*, Part II, p.69-80.

Burkhardt J. (2003). Réalité virtuelle et ergonomie : quelques apports réciproques. *Le travail humain*, vol(66) : 65-91.

Chandler T., Cordeil M., Czauderna T., Dwyer T., Glowacki J., Goncu C. and Wilson E. (2015). Immersive analytics. IEEE, *Proc. 2015 Big Data Visual Analytics (BDVA),* pp. 1-8.

Fuchs P. (1996). Les interfaces de la réalité virtuelle. Paris : Les Presses de l'École des Mines de Paris.

Leblanc J.-M. and Pérès M. (2011) Propositions de visualisations pour la navigation lexicométrique, *Document numérique*, pp.105-134 du vol(14) :105-134. http://dn.revuesonline.com/article.jsp?articleId=16064

Lisle L., Chen X., Gitre J. E., North C. and Bowman D. A. (2020). Evaluating the Benefits of the Immersive Space to Think. In IEEE, *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. p. 331-337.

Férey N., Gros P. E., Hérisson J. and Gherbi R. (2004). Exploration by visualization of numerical and textual genomic data. *Journal of Biological Physics and Chemistry*, vol.(4), 102-110.

Hassenzahl M. and Tractinsky N. (2006). User experience-a research agenda. *Behaviour & information technology*, vol.(25-2), 91-97.

Narayan M., Waugh L., Zhang X., Bafna P. and Bowman D. (2005). Quantifying the benefits of immersion for collaboration in virtual environments. Association for Computing Machinery, *Proc.s of the ACM symposium on Virtual reality software and technology (VRST '05)*, pp.78–81.

Pérès M., (2016). Analyser avec des outils descriptifs et statistiques des infographies d'infographies, *LCN - Les Cahiers du Numérique,* vol(4) : 65-91.

Schmalstieg, D., Fuhrmann A., Hesina G., Szalavári Z., Encarnaçao L. M., Gervautz M. and Purgathofer W. (2002). The studierstube augmented reality project. Presence: Teleoperators & Virtual Environments, 11(1), 33-54.

Shrinivasan Y. B. and van Wijk J. J. (2008). Supporting the analytical reasoning process in information visualization. *Proc. of the SIGCHI conference on human factors in computing systems*, pp. 1237-1246.
Skarbez R., Polys N. F., Ogle J. T., North C. and Bowman D. A. (2019). Immersive Analytics: Theory and Research Agenda. *Frontiers in Robotics and AI,* vol(6): 1-15.
Szalavári Z., Schmalstieg D., Fuhrmann A., and Gervautz M. (1998). "Studierstube": An environment for collaboration in augmented reality. *Virtual Reality*, vol(3-1), pp.37-48.
Whitlock M., Smart S. and Szafir D. A. (2020). Graphical perception for immersive analytics. In IEEE, *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. p. 616-625.
Wong P. C. and Bergeron R. D. (1994). 30 years of multidimensional multivariate visualization. *Scientific Visualization,* vol.(2), pp.3-33.
Yang, Y., Cordeil M., Beyer J., Dwyer T., Marriott K. and Pfister H. (2020). Embodied Navigation in Immersive Abstract Data Visualization: Is Overview+ Detail or Zooming Better for 3D Scatterplots?. *IEEE Transactions on Visualization and Computer Graphics*, vol.(27), no(2): 1214-1224.

Vocazione e professione: l'imperativo morale dell'infermiera italiana nel periodo fascista

Loredana Piervisani[1], Mariachiara Figura[1], Paola Arcadi[1], Angela Durante[2], Ercole Vellone[1], Rosaria Alvaro[1]

[1]Università degli Studi di Roma Tor Vergata – loredana.piervisani@gmail.com chiarafigura9@gmail.com paola.arcadi@gmail.com ercole.vellone@uniroma2.it rosaria.alvaro@gmail.com
[2]Universidad de la Rioja Spagna – angela.durante@unirioja.es

Abstract
For centuries, nursing was considered a charitable profession. With the healthcare reform, fascism declared to educate nurses to be better trained and prepared. Considering the fascist legislative policies that led to professional group's initial definition, the study investigates the possible transition of the professional vocation concept from a religious character to a secular one. The Automatic Textual Data Analysis was carried out on 111 articles from the only professional journal of the time. On the one hand, the manuscript shows a nursing identity characterized by ascetic and holiness traits; on the other hand, it recalls the concept of asceticism typical of the Calvinist doctrine of ethical and moral irreproachability of daily life through the practice of the profession. A stereotyped nursing idea reinforced by religious and gender dimensions emerged from the study. Since stereotypes still exist, the paper aims to raise thoughts about implementing interventions to support the recognition of an intellectual profession.

Keywords: fascism, Licensed Nurse, stereotypes, vocation, moral, secularization, religion

Riassunto
La professione infermieristica è stata per secoli concepita per il suo carattere caritatevole. Il fascismo con la riforma assistenziale dichiarò di formare infermiere più colte e preparate. A fronte delle politiche legislative fasciste che determinarono una prima definizione di gruppo professionale, il paper indaga la possibile transizione del concetto di vocazione professionale da carattere religioso a laico. L'Analisi Automatica dei Dati Testuali ha esaminato 111 articoli dell'unica rivista professionale del tempo. Lo studio da un lato mostra un'identità infermieristica dai tratti ascetici e di santità, dall'altro richiama al concetto di ascesi della dottrina calvinista d'ineccepibilità etica e morale del quotidiano mediante l'esercizio della professione. L'indagine mostra una figura infermieristica stereotipata, rafforzata dalla dimensione religiosa e di genere. Poiché gli

stereotipi sono ancora attuali, il paper vuol contribuire ad aprire una riflessione, per attuare interventi a sostegno del riconoscimento di una professione intellettuale.

Parole chiave: fascismo, Infermiera Diplomata, stereotipi, vocazione, morale, laicizzazione, religione

1. Introduzione

Nei paesi occidentali ed in particolare nel contesto italiano, la storia dell'assistenza infermieristica è stata fortemente condizionata dalla cultura cristiana (Manzoni, 2016). Fonti secondarie affermano che la "prima luce infermieristica" si colloca nel 1500 con l'istituzione delle Congregazioni religiose (Manzoni, 2001), che per secoli attribuirono al concetto di assistenza e di ospedale il significato di un complesso di carità.

Con la Riforma sanitaria-assistenziale di fine '800 si tentò la laicizzazione degli ospedali, fallita a causa della sempre maggiore incidenza del mondo cattolico nelle scelte politiche (Cosmacini, 2005; Preti, 1984), e la donna consacrata rappresentò l'idealtipo nel campo assistenziale (Manzoni, 2001). Nei primi anni del '900 le suore ospedaliere, nel contesto di forte progresso medico-scientifico, iniziarono ad apparire manchevoli per le loro capacità tecniche (Sironi, 2012). Di contro la figura infermieristica laica era fortemente contestata per la sua condotta morale (Bezze, 2013).

La successiva riforma del settore assistenziale avvenne nel 1925 dal Regime fascista. Il provvedimento istituì la formazione nelle Scuole Convitto consentita alle sole donne, destinata alle nuove Infermiere Diplomate o Professionali (ospedaliere), alle Assistenti Sanitarie Visitatrici (Asv, ovvero infermiere specializzate nell'area sociale, operative sul territorio), e alle Caposala, che avrebbero dovuto sostituire le "vecchie" infermiere (generiche di ambo i sessi).

Lo spirito della Riforma italiana nacque sulla scia di quella inglese della seconda metà dell'800, patrocinata da Florence Nightingale, nobildonna inglese, anglicana, che interpretava la vocazione infermieristica come una chiamata divina (Martínez Tovar et al., 2017). Secondo questa impostazione d'Oltremanica alle nuove infermiere italiane fu richiesto forte zelo religioso, castità, ed irreprensibilità morale così come negli ordini religiosi (Onofri, 1941).

Il fascismo veicolò un'immagine ambivalente di infermiera: da un lato la "signorina" che guadagna spazio sociale e lavorativo (Bartoloni, 2012), dall'altro una figura dai contorni ascetici, casta e martire, (Fiumi, 1993).

Lo studio si muove in un quadro di assenza di studi primari di analisi del concetto di vocazione professionale durante gli anni del Regime Fascista, al quale va riconosciuta una prima definizione e strutturazione di gruppo professionale (Piervisani et al. 2018, 2021).

A fronte di un periodo storico centrale nello sviluppo della categoria, si ritiene indispensabile indagare la possibile transizione del concetto di vocazione

professionale infermieristica, da carattere religioso a laico, ovvero il passaggio da "vocazione religiosa" ad un weberiano "impegno appassionato".

2. Materiali e Metodi

La ricerca analizza gli articoli dell'unica rivista di categoria del tempo "L'Infermiera Italiana" pubblicati dal 1939 al 1943, periodo di massima irreggimentazione del Paese, mediante l'Analisi Automatica dei Dati Testuali (AADT) con software IRaMuTeQ sviluppato da Pierre Ratinaud (2009). I dati vengono statisticamente trattati sia a livello lessicale che testuale.

3. Analisi dei dati

Il bilancio lessicale mostra un corpus costituito da 111 articoli, di medio-grande dimensionalità (88225 occorrenze). Il testo può definirsi sia ricco che ricercato lessicalmente (Type/Token Ratio:10,17%; hapax: 5,21% delle occorrenze) (Bolasco, 2013).

Si è scelto di effettuare l'analisi delle specificità sulle forme attive per le diverse annate della rivista. Per quanto attiene ai termini sovrautilizzati con indice >2, nell'anno 1939 sono sovra-utilizzate le occorrenze che affrontano aspetti di rigore igienico-sanitario di natura assistenziale. Nell'anno 1940 i discorsi sono caratterizzati da occorrenze che evidenziano il connubio della donna-infermiera dalle innate doti materne, e di una professione totalizzante. L'analisi delle concordanze mostrano che l'infermiera assume l'appellativo di "autorità" perché chiamata a svolgere sulla terra e con privilegio l'opera di Dio. Emergono inoltre paragoni tra l'agiografia delle sante e la vita delle infermiere. L'anno 1941 vede la presenza di specificità sovra-utilizzate profondamente radicate alla matrice spirituale da cui muove l'agire professionale. L'anno 1942 è caratterizzato da termini che richiamano al massimo controllo delle infermiere, a cui è richiesta massima rettitudine ed equilibrio morale. Nel 1943 si riscontrano specificità legate al ruolo e alle conoscenze tecniche dell'infermiera specializzata (Asv), e alla sua metodica e rigorosa opera di assistenza sul territorio.

L'analisi continua con la cluster analysis e con l'individuazione di cinque gruppi semantici trattati nel corpus (Fig.1). I segmenti classificati sono stati l'88,96% del totale. La classe 1 denominata "inclinazione religiosa dell'infermiera diplomata" (26,1% dei segmenti totali del testo). La classe 2 (16,6%) è contraddistinta dal tema "norme corporativiste di differenziazione delle Infermiere Diplomate", la classe 3 (21%) si riferisce al tema della "formazione" aspetto fondante la riforma assistenziale fascista. La classe 4 (12,2%) ingloba il tema "dogmi cattolici-professionali ed ideologia e retorica fascista". La classe 5 racchiude il tema del "contesto territoriale e assistenza infermieristica al nucleo famigliare".

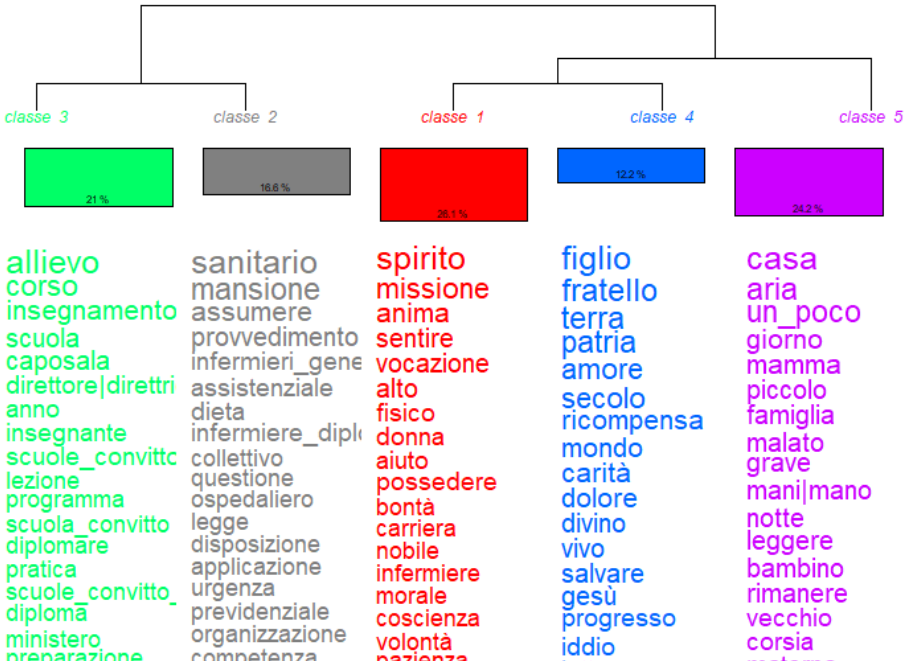
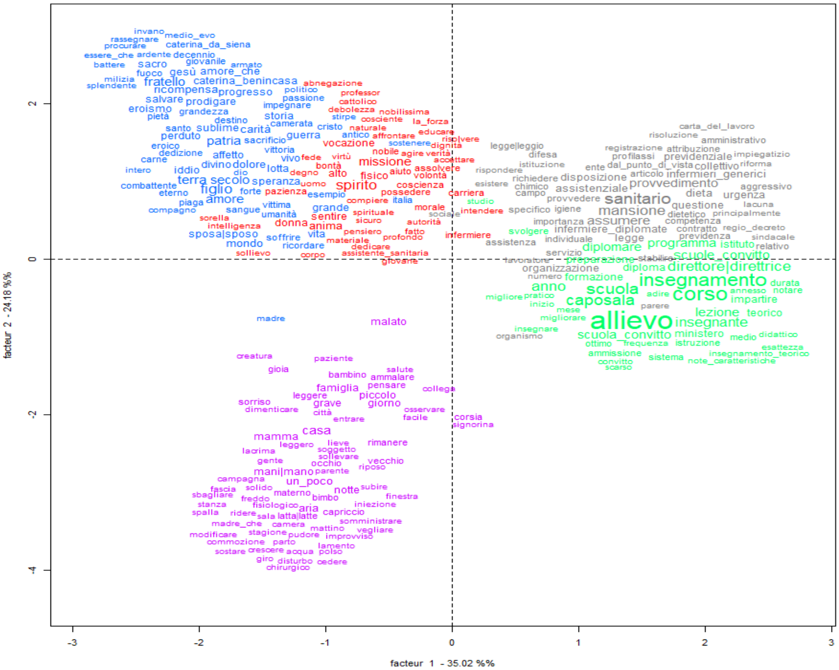


Figura 1. Cluster Analysis. Fonte: IRaMuTeQ, 2022.

L'analisi si conclude con la proiezione dei cluster nel piano fattoriale (Fig.2). Dall'interpretazione del grafico si risconta che la classe 5 "contesto territoriale e assistenza infermieristica al nucleo famigliare" è indipendente, discostandosi da tutte le altre nuvole. Vanno invece a legarsi a due a due la classe 3 nominata "formazione" con la classe 2 "norme corporativiste di differenziazione delle Infermiere Diplomate" e la classe 1 "inclinazione religiosa dell'Infermiera Diplomata" con la classe 4 "dogmi cristiani-professionali e ideologia/retorica fascista". Il fattore 1 (35,02% dell'inerzia totale) denominato "istituzionalizzazione percorso formativo" (nel semiasse positivo del fattore 1) vede la presenza delle seguenti occorrenze "mansione", "assumere", "Infermiere-Diplomate", "legge", "scuole-convitto", "corso", di istituzionalizzazione del percorso formativo necessario per l'accesso alla professione. Al lato opposto c'è la dimensione intima-religiosa e di genere nella cura" (semiasse negativo del fattore 1) troviamo "spirito" "vocazione" "donna" "anima" "pazienza" "madre" "malato", che frena lo sviluppo dell'identità infermieristica. Il secondo fattore (24,18% dell'inerzia totale) propone la dimensione "assistenza socio-familiare territoriale" (semiasse negativo del fattore 2), contraddistinto dai seguenti lemmi: "casa" "mamma" "famiglia" "somministrare" "latte". Tali occorrenze evidenziano la specializzazione professionale dell'Asv, che sembra prevalere

sulla visione religiosa della figura infermieristica, e che si scontra con l'identità veicolata dall'ideologia fascista. In contrapposizione c'è la dimensione nominata "fusione matrice cattolica con l'ambito professionale" (semi asse positivo del fattore 2) contraddistinta dalle occorrenze "eroismo", "Gesù", "sacro", "ricompensa", "Caterina_da_Siena", "passione".



Classe 1: rosso - Classe 2: grigio - Classe 3: verde - Classe 4: blu - Classe 5: viola

Figura 2. Analisi delle Corrispondenze (AC) con proiezione dei cluster sul piano fattoriale. Fonte: IRaMuTeQ, 2022.

4. Discussioni

L'indagine mostra quanto il topic in esame necessiti di un approccio interdisciplinare, laddove la disciplina infermieristica va ad integrarsi con quella storica e sociologica. La costante presenza della dimensione di genere ha concorso ad un'identità professionale stereotipata. Una rappresentazione preconcetta appesantita dalla forte retorica e propaganda fascista.

Il fascismo, l'ideologia su cui si fondava e le politiche attuate rappresentarono un'autoritaria volontà di cambiamento delle masse: delle idee degli individui, dei ruoli di genere, delle attività che ciascuna persona era chiamata a svolgere. Il Regime permeò integralmente nella vita privata del singolo (Dogliani, 2014).

Ciò trova pieno riscontro nei risultati di ricerca che evidenziano quanto l'attività professionale dell'infermiera donna fu totalizzante nella sua esperienza di vita.
In particolare, per quanto attiene al pressante controllo di genere il Ventennio Fascista ebbe un'organizzazione sociale fortemente impostata sulla separazione tra ruolo femminile e quello maschile, con una evidente subordinazione della donna all'uomo. Una divisione sociale che determinò l'inammissibilità da parte delle donne ad accedere alle occupazioni ritenute "naturalmente maschili" e di contro ad avere esclusività in altri campi lavorativi, come in quello assistenziale esercitato dalle infermiere professionali. Il corpo femminile venne nazionalizzato, per l'attribuzione della funzione riproduttiva e materna ideologicamente dovuta alla patria (Ropa and Venturoli, 2010).
Giovanni Gentile come citato in Ropa e Venturoli (2010, p. 141) ritenne che «la donna idealmente è madre prima di essere tale naturalmente madre per i figli, per i fratelli, per gli infermi per i piccoli affidati alla sua educazione». Gli stereotipi attribuiti alla donna furono sostanzialmente due ed in netta contrapposizione: «[…] una era la donna-crisi, cosmopolita, urbana, magra, isterica, decadente e sterile. L'altra era la donna-madre: patriottica, rurale, florida, forte, tranquilla e prolifica» (de Grazia 2007, p. 109), quest'ultima, identificata anche come "angelo del focolare", rappresentò l'ideale femminile fascista.
Il Regime impose l'obbligo morale di conformarsi ad un inflessibile codice di comportamento all'interno di ruoli precostituiti. La trasgressione di tali norme portava la donna all'emarginazione, all'invisibilità sociale, a ritenerla deviante per la sua psicologica instabilità, sino all'internamento nei manicomi (Valeriano and Di Sante, 2016). L'articolo delinea un parallelismo tra ruoli rigidi e predeterminati per la donna e gli schemi di una figura infermieristica concepita e normata in base al genere.

5. Conclusioni

Dall'AADT emerge una categoria infermieristica contraddistinta da un'identità imperniata su dogmi e precetti cattolici. Una rappresentazione lontana da quel processo di laicizzazione vocazionale che poteva aprire la strada alla professionalizzazione.
Una ideazione professionale che ebbe punti di contatto con la confessione anglicana, e con il concetto di "beruf" calvinista, di ineccepibilità etica e morale del quotidiano. Si era di fronte alla "quotidianizzazione dello straordinario" (Alagna, 2014), da cui nasce l'esigenza per il credente di avere una vita metodica e sistematica (Lamattina, 2017). Non vi furono spiragli legati al termine "beruf" che Max Weber definì un impegno appassionato verso la professione "intellettuale", una vocazione verso la scienza (Alagna, 2014). Piuttosto si trattò di un dovere assistenziale che Dio assegnò alla giovane donna, seppur laica, che richiedeva passione cristiana. L'analisi mostra un dato molto interessante, lo

scostamento del profilo delle Asv dal profilo sacro, in virtù della loro alta specializzazione. Concludendo si ritiene che l'indagine potrebbe contribuire ad una riflessione sulla categoria infermieristica, caratterizzata ancora oggi dalla presenza di stereotipi sociali che risentono di radici identitarie e culturali radicate nel concetto di religiosità (Piervisani et al., 2021). I risultati dello studio potrebbero far accrescere l'attenzione al focus di ricerca, e la necessità di interventi e misure atti a sostenere l'attuale riconoscimento di una professione intellettuale.

Bibliografia

Alagna M. D. (2014). *Immagini del mondo e forme della politica in Max Weber.* Università degli studi di Trento.
http://eprints-phd.biblio.unitn.it/1254/1/Alagna_Tesi_Dottorato.pdf
Bartoloni S. (2012). *Il fascismo e le donne nella "Rassegna Femminile Italiana" 1925-1930*. Roma: Biblink Editori.
Bezze S. (2013). Il contributo allo sviluppo della professione infermieristica delle Congregazioni Religiose del XIX secolo." *Professioni Infermieristiche,* vol. 66 (2): 75–84. doi: 10.7429/pi.2013.662075.
Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.
Cosmacini G. (2005). *Storia Della Medicina e Della Sanità in Italia*. Bari: Editori Laterza.
De Grazia V. (2007). Le Donne Nel Regime Fascista. 3rd ed. Venezia: Marsilio.
Dogliani P. (2014). Il Fascismo Degli Italiani: Una Storia Sociale. Novara: De Agostini.
Fiumi A. (1993). *Infermieri e Ospedale. Storia della professione infermieristica tra '800 e '900*. Verona: Nettuno.
Lamattina V. (2017). Capitalismo e ideologie della libertà. La libertà oltre il capitale. Università degli Studi di Salerno.
http://elea.unisa.it/jspui/bitstream/10556/3105/1/tesi%20di%20dottorato%20V.%20Lamattina.pdf
Manzoni E. (2001). Storia dell'assistenza. In Cantarelli, M. and Frati, L. (Eds.). *Storia e filosofia dell'assistenza infermieristica*, (pp. 2-90). Milano: Masson.
Manzoni E. (2016). *Le radici e le foglie: una visione storico-epistemologica della disciplina infermieristica*. 2nd ed. Torino: Casa Editrice Ambrosiana.
Martínez Tovar A., Rojas Cartillo X., Rosas Vargas L., Zárate Grajales R. A. and Flores Mora L. F. (2017). "Florence Nightingale, pionera en el modelo de gestiòn." *Revista CuidArte*, vol. 6 (11): 44-52. doi: 10.22201/fesi.23958979e.2017.6.11.69136.
Onofri F. (1941). Problemi spirituali dell'Asv nell'ora attuale. *L'Infermiera Italiana. Organo Ufficiale del sindacato nazionale fascista infermiere diplomate aderente alla Confederazione dei Professionisti e degli Artisti*, Biblioteca Nazionale Centrale di Firenze XIII.Re.167, vol. 7 (12): 4-7.
Piervisani L., Stasi S., Ruffinelli E., Vellone E., Rocco G. and Alvaro R. (2018). Storia e radici identitarie professionali: immagine dell'infermiera donna nel periodo fascista dal 1935 al 1938. *L'Infermiere*, vol. 55 (6): e134-49.

Piervisani L., Palombo A., Albanesi B., Rocco G., Stasi S., Vellone E. and Alvaro R. (2021). "The nurse in the mirror: image of the female nurse during the italian fascist period." Journal of Advanced Nursing, vol. 77 (2): 957-972. doi: 10.1111/jan.14679.
Preti D. (1984). La questione ospedaliera nell'Italia Fascista (1922-1940): Un aspetto della "modernizzazione Corporativa". In Preti, D. (Ed.). *Malattia e Medicina*, vol. 7, (pp. 335–87). Torino: Einaudi Editore.
Ratinaud P. (2009). IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. http://www.iramuteq.org.
Ropa R. and Venturoli C. (2010). *Donne e Lavoro: Un'identità difficile*. Bologna: Editrice Compositori.
Sironi C. (2012). *L'infermiere in Italia: storia di una professione*. Roma: Carocci Editore.
Valeriano A. and Di Sante C. (2016). *Mostra foto-documentaria "I Fiori Del Male". Donne in manicomio nel regime fascista*. Acquaviva Picena: Fast edit.

The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection

Bénédicte Pincemin[1], Serge Heiden[1], Franck Mazuet[2]

[1]University of Lyon – benedicte.pincemin@ens-lyon.fr slh@ens-lyon.fr
[2]University of Paris – fmazuet@gmail.com

Abstract
Active corpus provides the possibility to apply searching and statistical computing as if corpus were reduced to selected words, whereas full text still remains visible in context display. This is mainly implemented in paradigmatic processing, yet it may concern syntagmatic processing or text display too. Here we experiment active corpus in syntagmatic processing. A projection generates a new corpus, in which words are semantic tags that were automatically assigned in a first step to the original data. This new corpus makes it easy to explore tag sequences, with any generic textometric tool available, however sparse the original annotation may be. This methodological path was applied to film grammar analysis on 10,000 archival descriptions of news reports. 19 camera shot and angle types were ed through queries and tagged. This annotation became the lexicon of the projected corpus that was used to study shot sequences. The annotation and projection tools we have run are available as utilities in TXM open-source software and should usefully serve many research projects.

Keywords: textometry, active corpus, corpus annotation, corpus projection, film grammar, Les Actualités françaises newsreels, TXM software.

1. Definition and state of the art

1.1. What is active corpus?

We define *active corpus* as the subset of textual data that is taken into account in the textometric processing. It may be called *background* as well, as it provides the total amount of frequencies for contrastive analysis (Pincemin, 2004). This concept is useful to manage two different realities: the full source text and its relevant content for various processing.

Active corpus plays a role at several moments of textometric analysis. *(a)* When the text is *displayed*, parts of the text that do not belong to active corpus may be shown anyway (because the whole text matters in order to achieve the reading

and interpretation) but with a distinctive layout (such as a gray print), so as to know what is actually considered. *(b)* For *syntagmatic processing* dealing with the flow of the text and considering words as successive *tokens* (like phrase search or n-gram computing), active corpus is a means for passing over secondary items (insertions, expansions). *(c)* For *paradigmatic processing* listing words out of context and operating on *types* (that unify repeated occurrences), such as statistics on lexical tables, active corpus selects the focus (the items for which results are requested) and the overall set that stands for the frequency distribution (reference corpus and table margins).

1.2. Current implementations

Active corpus is commonly implemented for *paradigmatic processing*, through several ways to adjust the rows of the lexical tables submitted to statistical computing that can be:

- qualitative filtering: stop word list, part-of-speech selection, or analytic categories as assigned by an editable dictionary, like analytic keys in IRaMuTeQ that specify if a word is active, supplementary or excluded (Loubère and Ratinaud, 2014);
- quantitative threshold: either the *n* most frequent words, or words with a frequency higher than a given threshold;
- table advanced building: compiling several selection results, merging or discarding rows, like the *List* function in Hyperbase (Brunet, 2011).

Active corpus is then defined at the word-type level (in a dictionary, in indexes, in table rows) and is used in statistics computing. A preselection at the word-token level can also be combined (restriction to a subcorpus). Most frequently are the selected lexical rows both the statistical reference and the vocabulary under study — the two sets are identical. However, supplementary elements make it possible for a word to be considered in the analysis without weighing in the statistical reference. Conversely, one could display in the results only a subset of the words involved in the statistical calculus.

Concerning the active corpus in *text display*, this is rarely available in text analysis tools. Such a feature is more suitable to TEI-based tools that encode and render complex text internal structures, like TEITOK (Janssen, 2016).

As for active corpus in *syntagmatic processing*, a current illustration may be provided by the possibility, in a concordance view, to consider only some word positions when sorting contexts: for instance, the 2L position meaning “the second word to the left of the target word” (see AntConc for an example of implementation (Anthony, 2022)). Nevertheless, active corpus appears still underused in syntagmatic processing.

1.3. Active corpus in TXM software

TXM implements the usual means to refine a lexical table to fit any active corpus (*paradigmatic processing*): token preselection (subcorpus), qualitative filtering (CQP search engine for rows selection), quantitative thresholds, operations on table rows. Moreover, the lexical table margins can be set either to the current row selection, or to the entire corpus or subcorpus, which allows focusing results on a subset of words while keeping a larger corpus as reference for frequencies. The direct access to specificity score computation through its four parameters (PlotSpecif utility, Fig. 4 below) provides the most accurate settings for the active corpus and the pattern to be studied, for any particular case of specificity evaluation.

TXM also provides means to deal with active corpus when *displaying the text*: in the XML-TEI Zero import process (Heiden, 2020: §4.9.5), the textual planes setting labelled "Out of text to edit" lists XML elements whose textual data content will be displayed without being indexed (these words are neither searched nor counted). And a CSS stylesheet can be edited to distinguish the layout for these elements if necessary. An example is given with red section titles in the text displayed in Figure 1.

A recent utility (WordProperty2Word) carries out a projection of a word annotation that creates a corpus whose words are selected annotation values. In doing so, it can retain a subset of original tokens and delete others. We argue this can be a pragmatic way to implement active corpus for subsequent *syntagmatic* analyses, as illustrated in the following section.

2. A use case scenario

2.1. Research context and addressed question

The ANTRACT research project (Carrive *et al*., 2021) studies the audiovisual corpus of *Les Actualités françaises*, composed of about 10,000 news reports that were broadcast as weekly editions in French movie theaters between 1945 and 1969. One of the fields of investigation is an historical description of the production process and its impact on the newsreel's features. From this perspective, the question was raised to capture data about the film grammar of these newsreels. That is, what kind of camera shots and angles do the newsreels actually use, and whether recurring patterns of shots and angle types contribute to film standardization.

We considered that such information might be obtained from the newsreel descriptive forms. In these text documents, the *Sequences* section is a shot-by-shot description of the film footage (see example in Figure 1 and Appendix). We must keep in mind that this is secondary data, as it relies on librarian mediation, which can vary from one librarian to another, and which cannot be totally exhaustive. Nevertheless, we thought these professional descriptions were worth

mining and might give insights on film grammar.

2.2. Analysis roadmap within TXM

A TXM formatted AF-NOTICES corpus has been created from the descriptive forms (Pincemin *et al*., 2020). Its size amounts to 2 million words, among which 38% hapax (more information in Appendix). It is a structured corpus, in which *Sequences* sections can be targeted. The idea is to set a typology of shot scales: Close-Up, Medium Shot, etc. For each shot type, CQL search engine queries are elaborated to match the various wording librarians happen to use to mention the shot. An automatic annotation utility applies the queries to the corpus and adds a normalized encoding for shots, available as a new analytic lexical property. Then, we project the corpus along this property, so as to derive a new corpus that retains only sequences of shot types. In this new corpus, shot pattern analysis is no longer hindered, neither by wording variations nor by lexical insertions that break up shot strings in the original text.

2.3. Annotation step

Several annotation modes are available in TXM platform (Heiden, 2018). We opted here for a lexical annotation, assigning a property to a word, for two reasons. First, this annotation is indexed in the CQP model so that it benefits from all TXM usual functionalities. Secondly, a lexical annotation fits the targeted one-shot/one-unit model.

We also chose to systematize the description as every association between a term and a shot value was formally encoded through a CQL query. In addition to the consistency and clarity of the model, this procedure is compatible with corpus updates, and even with description updates (categories can be refined while getting more experienced with the data).

Our model lists 19 camera shot and angle types, plus one miscellaneous type (“divers plans”), and one off-topic type that is used to manage polysemy and mark “plan” or “vue” occurrences that do not denote a shot or an angle. For our needs, 60 queries were designed (see Appendix). We benefited from the CQLList2WordProperties utility to automatically apply these 60 queries to the corpus and encode the new “plan” word property, so that annotating was systematic and fast.

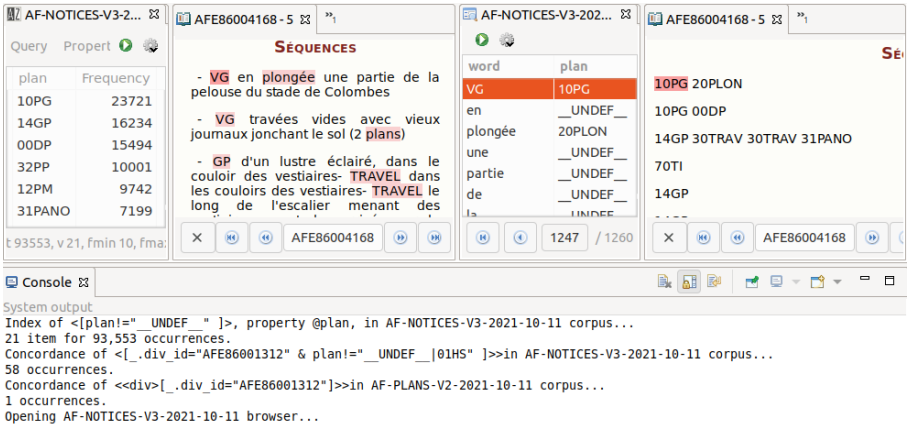


Figure 1. The Les Actualités françaises *corpus in TXM, with shot type annotations. From left to right: (i) annotated shot tags and their frequencies in the corpus; (ii) view of a documentary description with all annotated words highlighted; (iii) view of the shot tag assigned to each word of this text; (iv) same text in its projected version. Below: log of executed commands.*

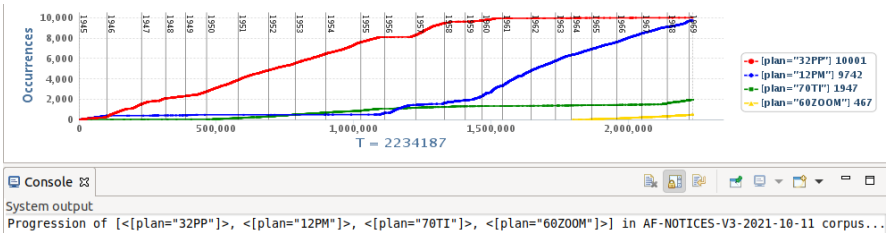


Figure 2. Progression chart showing the diachronic evolution of 4 shot types. By decreasing frequency: Handheld Shot (red), Medium Shot (blue), Title & Credits (green), Zoom In & Out (yellow)

Figure 1 shows a summary of the resulting annotation (a developed example is provided in the Appendix). The shot that is the most frequently noted in descriptions is the Extreme Wide Shot (10PG) with 23,721 occurrences, which represents one quarter of all annotated shot mentions (93,553 annotations) in a 2-million-word corpus. For instance, one can examine how shot mentions are distributed along time (Figure 2) and note that Handheld Shots seems to have been replaced by Medium Shots. Another pattern involves the use of Zoom lenses during newsreel shootings. The appearance of this type of camera optics in the progression chart signals its adoption by *Les Actualités françaises* cameramen in 1964 and its increasing use among the company crews from then on.

2.4. Projection step

The projection (WordProperty2Word) of the annotated corpus along its shot-type annotation ("plan" word property) creates a new corpus whose words are the sequence of shot types (see Figure 1, right panel). Thus, shot mentions are normalized, and other words are ignored. This new corpus is 91,898-word long. It can be explored with all usual textometric tools.

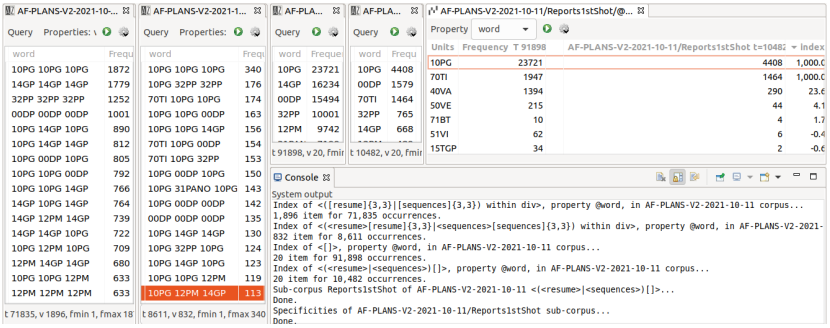


Figure 3. Study of the film grammar in Les Actualités françaises.
From left to right: (i) frequency list of 3-shot patterns and (ii) same thing for initial 3-shot patterns only; (iii) frequency list of shots and (iv) same thing for first shot in a report only; (iv) specificity score of shot types for the first position in reports. Bottom right panel: log of executed commands.

An illustration of the usefulness of this new representation is the study of the patterns of shots beginning a report (Fig. 3 and Appendix): we can search and count for 3-shot patterns, whether they begin the report (ii) or occur anywhere within a report (i); we can formulate a basic query matching the first shot in a report (iv), and measure specificities of shot types in this leading position (v). Specificity can be computed for patterns too (Figure 4).

We then were able to conduct a comprehensive analysis (see Appendix) of how scale evolves in the very first shots. 9,551 reports have got at least two shots mentioned in their text description. Among these, 5,582 reports (58 %) have a second shot type that is at a closer scale than the first one, which we call a funnel pattern. 2,065 other reports (22 %) use a unique and constant type of shot in their description. So all other patterns (either a second shot type wider than the first one, or a funnel pattern occurring later than in the very first shots) happen in the few remaining 1,904 reports (20 %). We interpret this prominence of the funnel shot pattern as a kind of standardization of film grammar. This approach has a dual goal. It offers the public a didactical presentation of the news based upon a simple editing where the gradual narrowing of successive shot frames enables a fluid visual exposition of any given subject. It also uses the

standardization and reiteration of the *ad hoc* sequences to accelerate the shooting and editing process, thus providing an economical production method to the newsreel companies.

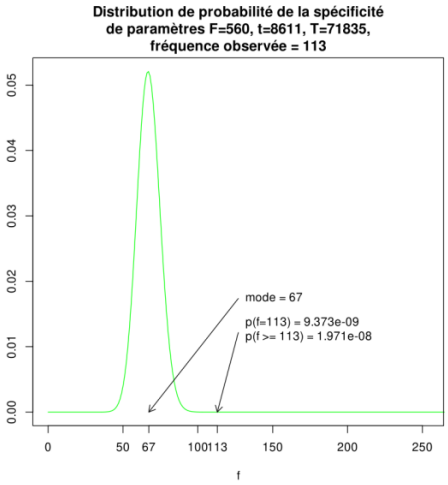


Figure 4. Specificity result +7.7 for the funnel pattern10GP 12PM 14GP (Extreme Wide Shot, Medium Shot, close-Up) as first shot sequence in a report (TXM PlotSpecif utility output).

A drawback of this projection-based approach for active corpus is some loss of context, since original words are no longer present. However, one can work simultaneously with both corpora (the one before and the one after projection) and view both representations of the text (see example Fig. 1 with panels *(ii)* and *(iv)*). Textual structures (sections, paragraphs) and their metadata are preserved, and each shot-type word in the projected corpus has got a property that records the lexical form of the original tagged word in the shot mention.

3. Summary and prospects

Active corpus is a critical parameter in textometric analysis as it allows finetuning of textual data for searching domain and for statistical computing. On the one hand, active corpus is quite fully implemented in textometric software as regards *paradigmatic* processing (on lexical tables); on the other hand, it is much less advanced concerning text *display* and *syntagmatic* processing. For this last case, we devised and tested a new analysis scenario based on utilities that were recently released in TXM open-source software. First step is annotating through queries, which enriches the corpus with consistent and relevant analytic categories, even on evolving data with periodical releases. Second step is a projection that derives a new corpus focused on the chosen analytic categories. Both corpora are useful to the analysis: the original one provides the lexical and

textual contexts when looking back at text content; and the projected one is efficient to address textometric questions about analytic patterns. Such a strategy is worth disseminating since it provides an effective solution for many similar research scenarios and needs. A more powerful implementation of active corpus for syntagmatic processing may still be achieved with a full integration of analytic representation and textual context into a unique corpus.

This research was supported by the ANTRACT ANR project (ANR-17-CE38-0010).

References

Anthony L. (2022). *AntConc (Windows, MacOS, Linux) Build 4.0.5. Help file*. Waseda University.

Brunet É. (2011). *Hyperbase, Manuel de référence*. Université de Nice.

Carrive J., Beloued A., Goetschel P., Heiden S., Laurent A., Lisena P., Mazuet F., Meignier S., Pincemin B., Poels G., and Troncy R. (2021). Transdiciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quarterly*, 15 (1).

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al., ed., *Proc. PACLIC 2010*. Waseda, 389-398.

Heiden S. (2018). Annotation-based Digital Text Corpora Analysis within the TXM Platform. In Iezzi D. F. et al. editors, *Proc. of JADT 2018*. Roma: UniversItalia, pp.367-374.

Heiden S. (2020). *Manuel de TXM, v. 0.8*. ENS de Lyon. https://pages.textometrie.org/txm-manual/

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In Calzolari, N. *et al*., editors, *Proc. of LREC'2016*. Portorož, Slovenia, pp. 4037-4043.

Loubère L. and Ratinaud P. (2014). *Documentation IRaMuTeQ 0.6 alpha 3 v. 0.1*. Univ. Toulouse.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In Purnelle G. et al., editors, *Proc. of JADT 2004*. Presses universitaires de Louvain, vol. II, pp. 865-873.

Pincemin B., Heiden S., and Decorde M. (2020). Textometry on Audiovisual Corpora: Experiments with TXM software. In Ratinaud P. and Marchand P., editors, *Proc. of JADT 2020*. Univ. Toulouse 3.

Appendix

Full resolution figures and supplementary materials are available in the HAL-SHS archive: https://halshs.archives-ouvertes.fr/halshs-03667319

Stylométrie des *Elemens de la grammaire Françoise* (1780) ou comment Lhomond a supplanté ses prédécesseurs

Sophie Piron

Université du Québec à Montréal – piron.sophie@uqam.ca

Abstract
This study uses descriptive clues to establish the stylistic specificity of Lhomond's Elements of the French Grammar, published in 1780. This elementary grammar was immediately successful and initiated the current of French school grammar. It supplanted the two most important grammars that preceded it within the elementary grammars, namely those of Restaut (1732) and Wailly (1759). However, the book is not theoretically innovative. The hypothesis developed here is that the success of this grammar owes a lot to its stylometric characteristics. These are compared with those of the grammars of Restaut and Wailly.

Keywords: stylometry, French grammar, 18th century, Lhomond, Restaut, Wailly.

Résumé
La présente étude exploite des indices descriptifs pour établir la spécificité stylistique des *Elémens de la grammaire françoise* de Lhomond, parus en 1780. Cette petite grammaire a connu un immense succès et initie le courant de la grammaire scolaire du français. Elle a supplanté les deux grammaires phares qui l'ont précédée dans le courant des grammaires élémentaires, soit celles de Restaut (1732) et de Wailly (1759). L'ouvrage n'est pourtant pas novateur sur le plan théorique. L'hypothèse développée ici est que le succès de cette grammaire doit beaucoup à ses caractéristiques stylométriques. Celles-ci sont comparées à celles des grammaires de Restaut et de Wailly.

Mots clés: stylométrie, grammaire française, 18e siècle, Lhomond, Restaut, Wailly.

1. Introduction

La présente étude[1] s'inscrit dans une recherche plus vaste qui a pour objectif de proposer une édition critique des *Elemens de la grammaire françoise* (EGF) de

[1] Nous tenons à remercier chaleureusement l'évaluateur 2 pour ses remarques éclairantes.

Charles-François Lhomond (1727-1794), parus en 1780, à Paris. L'ouvrage a connu le succès, du vivant même de l'auteur puisqu'il a fait l'objet de sept éditions entre 1780 et 1790. De nombreuses éditions posthumes ont ensuite été produites jusqu'à la fin du 19e siècle. L'ouvrage a incontestablement marqué l'histoire de la grammaire scolaire du français et est reconnu à ce titre.

La date clé dans l'histoire de la production grammaticale à usage scolaire au XVIIIe siècle n'est pas le Restaut de 1732, qui distingue les deux orthographes, ni le Wailly de 1754 qui abandonne la déclinaison, mais le petit Lhomond de 1780. (Chervel 2006 : 220)

L'origine de cette postérité ne s'explique pas par la profondeur de l'exposé. Lhomond ne propose, en effet, aucun classement novateur, aucune avancée grammaticale. Son seul apport d'ordre théorique est d'avoir entériné la scission du substantif et de l'adjectif en deux parties du discours autonomes, mais il n'est pas celui qui est à l'origine de cette innovation, qui a été proposée par deux grammairiens notoires, Girard (1747) et Beauzée (1767). Peu d'auteurs à l'époque l'avaient encore acceptée, Lhomond l'a fait passer dans la doxa grammaticale. Le véritable apport de Lhomond se situe donc ailleurs. Dans la forme adoptée, d'abord : le style est simple et concis. Dans le contenu proposé, ensuite : la matière est condensée et élaguée. En effet, les *Elemens* tiennent en un peu moins de 100 pages. Ce « degré zéro du manuel » de grammaire (Chervel 2006 : 220) est ce qui a fortement contribué à son succès dans un monde scolaire balbutiant, où élèves patoisants et maîtres mal formés avaient besoin de simplicité.

Chervel (1977) fait de Lhomond le « continuateur » de Restaut et de Wailly dans la mesure où ces deux grammairiens ont offert, respectivement en 1732 et en 1759, des abrégés célèbres qui sont destinés aux débutants, le même public auquel s'adresse Lhomond. Les œuvres de Restaut et de Wailly constituent des publications phares du 18e siècle (Fournier 1998, Julien 1998). Les deux auteurs sont régulièrement cités à l'époque et leurs ouvrages connaissent une diffusion importante. L'*Abrégé* de Restaut est utilisé comme manuel dans les collèges jusqu'à la fin du 18e siècle, et l'A*brégé* de Wailly compte douze éditions jusqu'en 1804. Or Lhomond vient concurrencer ces deux ténors de la grammaire élémentaire et finit par sonner le glas de leur succès puisque les *Elemens* de 1780 sont presque immédiatement adoptés dans les collèges et les écoles royales militaires (Chervel 2006) et primés en 1796 au concours pour les livres élémentaires. Les textes officiels de l'instruction publique française y feront ensuite périodiquement référence au cours du 19e siècle (Chervel 1992).

L'objectif de l'étude proposée ici est de trouver en quoi Lhomond a réussi à supplanter les *Abrégés* de Restaut et de Wailly. Pour ce faire, les sections suivantes détaillent d'abord la méthodologie utilisée et abordent ensuite la répartition du contenu des trois ouvrages, la longueur des phrases, la richesse du vocabulaire et enfin la complexité syntaxique.

2. Méthodologie

Dans le domaine de l'histoire de la grammaire, les critères méthodologiques portent avant tout sur le choix de l'édition de référence. L'édition choisie pour les *Elemens* est celle de 1790, la dernière du vivant de l'auteur. Elle ne présente, en fait, aucune modification de contenu par rapport à celle de 1780, seulement des améliorations typographiques. Comme le marché du livre est moins réactif qu'à notre époque, nous avons choisi, pour la comparaison avec Restaut et Wailly, de prendre l'édition de 1745 pour l'*Abrégé* de Restaut et l'édition originale de 1759 pour l'*Abrégé* de Wailly. Nous aurons ainsi une image de ce qui plaît au milieu du 18e siècle en matière de livre de grammaire élémentaire.

Les trois documents d'époque ont d'abord été transcrits en format Word. Le texte de Wailly présentait une quantité impressionnante d'abréviations, qu'il a fallu résoudre. Les trois textes ont ensuite été segmentés manuellement en unités de natures diverses et de longueurs variées, mais toujours exclusives (chaque mot se trouvant dans une unité et une seule) : 1) le texte proprement dit, découpé en phrases graphiques (c'est-à-dire depuis une majuscule initiale jusqu'à une ponctuation finale) ; 2) le découpage, ce qui comprend les titres sous forme de texte, mais aussi des numérotations et indications diverses que l'on peut trouver en début de paragraphe en certaines occasions ; 3) les exemples autonomes (c'est-à-dire ceux qui ne sont pas intégrés à l'intérieur d'une phrase explicative, ils peuvent être détachés typographiquement ou apparaître à l'intérieur d'un paragraphe, entre deux phrases de texte ; ils sont surtout clos par une ponctuation finale qui leur est propre) ; 4) les tableaux (paradigmes des pronoms, des articles et de la conjugaison). L'analyse syntaxique a été réalisée grâce au logiciel Antidote.

3. La répartition du contenu des ouvrages

À première vue, les ouvrages présentent déjà quelques différences : les EGF comptent le nombre de pages le plus réduit et le texte y est relativement aéré. C'est par conséquent chez Lhomond que le nombre brut d'unités est le moins élevé. Il est de 917. La comparaison montre, chez Lhomond, une diminution du contenu textuel, une augmentation des tableaux et une exploitation intermédiaire des exemples et du titrage. Toutefois, si l'on envisage le corpus Restaut un peu différemment, certaines tendances apparaissent plus clairement. En effet, cette grammaire se différencie nettement des deux autres sur le plan rédactionnel parce qu'elle est rédigée sous forme de questions et de réponses. Il s'agit d'une pratique d'écriture classique, dite *par demande* (D) *et réponse* (R). Restaut fait partie des derniers grammairiens à l'exploiter. Dans les faits, le texte se présente comme suit :

D. Qu'est-ce qu'un nom ?
R. C'est un mot qui sert à nommer, ou à qualifier quelque chose.

D. Combien y a-t-il de sortes de noms ?
R. Deux ; le Nom substantif et le Nom adjectif. (Restaut 1745 : 15)

La version A du corpus Restaut comptabilise les 346 abréviations « D. » et autant de « R. » dans les unités de découpage. Or ces indicateurs fournissent une information dialogique non seulement répétitive et prévisible, mais surtout fondamentalement différente du découpage au moyen de titres et de numérotations qui procèdent, pour leur part, à une organisation thématique du contenu. Dès lors, la version B du corpus laisse délibérément de côté ces unités D et R.

Tableau 1. Les unités constitutives des corpus

	Restaut (1745) : A		Restaut (1745) : B		Wailly (1759)		Lhomond (1790)	
Unités de texte	931	51 %	931	83 %	1046	53,5 %	405	44 %
Unités d'exemples	26	1 %	26	2 %	421	21,5 %	127	14 %
Unités de tableaux	41	2 %	41	4 %	6	0,5 %	58	6 %
Unités de découpage	816	45 %	124	11 %	479	24,5 %	327	36 %
Total des unités	1814	100 %	1122	100 %	1952	100 %	917	100 %
Total des pages[2]	142 pages				143 pages		87 pages	

La comparaison avec la version B du corpus Restaut permet, cette fois, de constater deux évolutions importantes. La première est une tendance à diminuer le contenu textuel des ouvrages. Avec 44 % d'unités de texte, Lhomond a presque divisé par deux ce que proposait Restaut. La seconde évolution notable est une nette augmentation du découpage organisationnel thématique (le titrage au moyen de titres et de numérotations), permettant ainsi une lecture plus aisée, plus balisée, plus propice à une exploitation pédagogique. Ces unités organisationnelles sont ainsi passées de 11 % chez Restaut, à 24,5 % chez Wailly et à 36 % chez Lhomond.

Lhomond propose aussi plus d'exemples indépendants que Restaut, mais moins que Wailly, ce qui confère à la portion textuelle de l'ouvrage (texte et exemples) un poids qui reste en deçà des 60 %, contrairement à Restaut (85 %) et à Wailly (75 %). Lhomond est également celui qui fournit le plus de tableaux. La lecture est ainsi facilitée par la visualisation des paradigmes de formes. Lhomond a

[2] Il s'agit du nombre de pages imprimées dans l'édition consultée, pour le contenu de la grammaire, ce qui exclut la page de titre et les éventuels index, tables des matières, préfaces, avertissements et colophons.

accentué le choix de Restaut et s'est nettement distingué de l'emploi plus que parcimonieux que Wailly fait de cet outil de présentation pédagogique qu'est le tableau.

4. La longueur des phrases

La longueur des phrases constituant une caractéristique stylistique importante, nous avons comparé sous cet angle trois parties du corpus, en l'occurrence les unités de texte des trois ouvrages. Nous avons introduit un critère de classement dans le corpus Restaut en identifiant les unités selon qu'elles correspondaient à des demandes (D) ou à des réponses (R).

Il ressort de ces données que les textes sont relativement semblables dans la longueur maximale des phrases, de 85 à 94 mots. Ces ouvrages élémentaires du 18e siècle se caractérisent tous par des phrases plutôt longues, même dans les longueurs moyennes, qui oscillent de 18 à 24 mots. On relève toutefois que Lhomond rédige des phrases plus longues que ses prédécesseurs puisqu'elles montent en moyenne à 24 mots par phrase, face à 18 ou 19 mots chez Restaut et Wailly. Cette caractéristique rédactionnelle des *Elemens* correspond, en fait, exactement à la longueur moyenne que présente le sous-corpus des réponses dans Restaut, c'est-à-dire la partie plus proprement explicative, celle dont les caractéristiques sont le plus directement comparables avec les textes que rédigent Wailly et Lhomond. Il semble donc que Lhomond soit resté très classique sur ce point et n'ait pas proposé un texte stylistiquement simplifié tandis que Wailly a opté pour des phrases un peu plus courtes.

Tableau 2. La longueur des phrases dans les unités de texte

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Longueur moyenne	18 mots/phrase (D & R) 9 mots/phrase (D) ; 24 mots/phrase (R)	19 mots/phrase	24 mots/phrase
Longueur maximale	92 mots/phrase (D & R) 29 mots/phrase (D) ; 92 mots/phrase (R)	85 mots/phrase	94 mots/phrase

Il convient également d'aborder les exemples indépendants que contiennent les trois ouvrages. Les auteurs les exploitent différemment. Wailly propose notamment mais pas exclusivement des citations d'auteurs, forcément plus longues que les exemples forgés que fournissent Restaut et Lhomond. Des exemples plus courts facilitent la compréhension des contenus grammaticaux exemplifiés. Lhomond et Restaut sont donc ici plus pédagogues que Wailly, mais Lhomond l'est encore davantage puisque la longueur maximale de ses exemples est la plus basse des trois corpus (31 mots contre 66 chez Wailly, et 96 chez Restaut).

Tableau 3. La longueur des phrases dans les unités d'exemples

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Longueur moyenne	7 mots/phrase	19 mots/phrase	7 mots/phrase
Longueur maximale	96 mots/phrase	66 mots/phrase	31 mots/phrase

5. La richesse du vocabulaire

Pour analyser la richesse du vocabulaire, les unités de texte des trois corpus ont été récupérées et ensuite lemmatisées grâce au correcteur Antidote. Voici les gammes de fréquences obtenues.

Tableau 4. La fréquence des formes dans les unités de texte

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Formes de fréquence 1	350 (37 % des formes)	976 (48 % des formes)	428 (45 % des formes)
Fréquences maximales	933 (*de*), 455 (*être*)	936 (*de*), 592 (*être*)	372 (*de*), 233 (*être*)
Nombre de formes	954	2016	954
Nombre d'occurrences	12620	16992	6910
Richesse du vocabulaire[3]	0,08	0,12	0,14

On peut dresser le constat que les hapax sont proportionnellement moins importants chez Restaut, ce qui engage une exploitation plus pauvre et répétitive du lexique que chez Wailly et Lhomond, lesquels présentent des valeurs usuelles à cet égard. Les fréquences maximales relevées touchent les mêmes formes dans les trois corpus, la préposition *de* et le verbe *être.* Elles présentent des quantités très semblables entre Restaut et Wailly alors que ce dernier fait usage de plus d'occurrences. Restaut se caractérise donc par une répétitivité plus grande. Les indices de richesse du vocabulaire confirment que le texte de Restaut présente la richesse la plus basse. La richesse du vocabulaire chez Lhomond engage un important bassin de formes en partage avec Restaut (59 %, 566 des 954 formes) et davantage encore avec Wailly (73 %, 694 des 954 formes). Les trois auteurs ont 493 formes en commun, soit 52 % des formes chez Lhomond. On peut donc supposer que ce dernier exploite un lexique simple et fréquent.

[3] Nous utilisons la formule basique selon laquelle la richesse du vocabulaire est calculée en divisant le nombre de formes par le nombre d'occurrences. Comme cet indice présente « l'inconvénient de trop dépendre de la longueur des textes » (Lebart et Salem 1994 : 248), une phase ultérieure de la recherche exploitera d'autres indices moins sensibles à la longueur, de manière à contrer ce défaut de comparativité qui défavorise fort probablement Wailly dans l'analyse actuelle.

6. La complexité syntaxique des phrases

Nous proposons trois indicateurs pour évaluer la complexité syntaxique. Ils imposent une catégorisation syntaxique préalable, ce qui a été réalisé au moyen d'Antidote. Le premier indicateur rapporte la présence d'adverbes (ADV) modifieurs lourds en *-ment.* Proto-typiquement, ils modifient des verbes et complexifient, par conséquent, la structure syntaxique.

Tableau 5. Indicateur de complexité syntaxique: les modifieurs adverbiaux en -ment

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Nombre de formes ADV-MENT	26	38	18
Nombre d'occurrences ADV-MENT (a)	44	73	37
Nombre d'occurrences total (b)	12620	16992	6910
Indice de complexité syntaxique (a/b)	0,00349	0,00430	0,00535

La complexité la plus faible se trouve chez Restaut, où elle est d'ailleurs souvent constituée d'hapax (*absolument, nécessairement*, *autrement, tellement*). La répétition y est concentrée sur un adverbe (*ordinairement*). C'est ce même adverbe qui est le plus répété chez les trois auteurs. Lhomond est celui qui réutilise le plus les adverbes (*seulement, rarement, entièrement*) et montre proportionnellement un indice de complexité légèrement plus élevé.

Le deuxième indicateur consigne la présence de ligateurs (LIG), c'est-à-dire des mots permettant de construire des syntagmes complexes ou des subordonnées. Il s'agit des prépositions, des coordonnants et des subordonnants. L'avantage de l'utilisation d'Antidote est qu'il repère les segments (tels que *afin de, près de, quant à,* etc.) et les traite comme des unités. Malheureusement, certains ligateurs ambigus étant catégorisés comme pronoms par Antidote, il a fallu les exclure ici. En effet, des formes comme *qui* ou *duquel* peuvent être de simples pronoms (*qui est là ?*) ou des subordonnants (*la personne qui écrit*). Or seuls les subordonnants sont des indicateurs de complexité syntaxique puisqu'ils introduisent des subordonnées.

Tableau 6. Indicateur de complexité syntaxique: les ligateurs

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Nombre de formes LIG	91	125	71
Nombre d'occurrences LIG (a) [F = 9361]	3425	4362	1574
Nombre d'occurrences total (b) [T = 36522]	12620	16992	6910
Indice de complexité syntaxique (a/b)	0,271	0,257	0,228
S+	6,0	0,4	-9,2

Il ressort vraisemblablement de ce tableau que Restaut sur-emploie les ligateurs alors que Lhomond les sous-emploie. Wailly se situe, pour sa part, dans une zone d'emploi non marquée.
Le troisième indicateur concerne la présence de pronoms anaphoriques (*celui-ci, plusieurs*, *lui*, *en, y,* etc.), c'est-à-dire des mots pour lesquels, à la lecture, un calcul sémantique doit être mené pour trouver le référent dans les segments textuels qui précèdent.

Tableau 7. Indicateur de complexité syntaxique : les pronoms anaphoriques

	Restaut (1745)	Wailly (1759)	Lhomond (1790)
Nombre de formes PRO	26	32	26
Nombre d'occurrences PRO (a)	978	976	438
Nombre d'occurrences total (b)	12620	16992	6910
Indice de complexité syntaxique (a/b)	0,077	0,057	0,063

Lhomond et Restaut font usage d'un bassin de formes pronominales un peu plus réduit que celui de Wailly. Ce dernier est pourtant celui qui présente l'indice de complexité le plus bas. Restaut, par son procédé rédactionnel *demande/réponse,* a davantage recours aux pronoms puisqu'ils peuvent être utilisés dans la réponse à une question. Lhomond présente sur ce critère un indice intermédiaire, mais qui reste relativement proche de Wailly.

7. Conclusion

L'analyse stylométrique qui vient d'être menée a permis de confirmer le ressenti que l'on a à la lecture des grammaires de Restaut, de Wailly et de Lhomond. Les *Elémens* constituent le texte le plus abordable de par son contenu et sa forme. Pour le contenu, Lhomond a réduit la proportion du texte et des exemples. Il a aussi fait basculer de nombreuses informations dans des tableaux. Pour la forme, il a facilité le parcours de lecture par un titrage qui balise les contenus. Tout en proposant des exemples très courts, il a opté pour des phrases longues dans la partie descriptive de l'ouvrage. En cela, il adopte un style rédactionnel classique, que l'on voit également transparaître dans la richesse du vocabulaire. Sur ce point, le texte présente des valeurs usuelles et se démarque de la simplicité de Restaut. Lhomond puise toutefois majoritairement dans un stock lexical commun, dans la mesure où la moitié des formes présentes dans sa grammaire se trouvent aussi exploitées par Restaut et les trois quarts par Wailly. Malgré des phrases longues et un vocabulaire plutôt développé (bien qu'issu d'un bassin usuel), le style lhomondien se caractérise par une complexité syntaxique qui ne se démarque pas substantiellement. Face aux trois critères de complexité,

Lhomond se situe tantôt en deçà des valeurs usuelles (les ligateurs), tantôt dans les valeurs communes (les pronoms anaphoriques), tantôt enfin un peu au-dessus des autres scores (adverbes en -*ment*). En résumé, Lhomond est classique, mais plutôt simple. Autrement dit, il simplifie sans briser les canons stylistiques. C'est ce qui lui a permis de supplanter ses prédécesseurs.

Bibliographie

Antidote 11. Druide informatique.
Chervel A. (1992). *L'enseignement du français à l'école primaire. Textes officiels. INRP/Economica.*
Chervel A. (2006). *Histoire de l'enseignement du français du XVII^e^ au XX^e^ siècle*. Retz.
Fournier J.-M. (1998). Restaut. *Corpus CTLF* [En ligne : http://ctlf.ens-lyon.fr/n_fiche.asp?n=316]
Julien J. (1998). Wailly. Corpus CTLF [En ligne : http://ctlf.ens-lyon.fr/n_fiche.asp?n=289]
Lebart L. et Salem A. (1994). *Statistique textuelle.* Dunod.
Lhomond Ch.-F. (1790 [1780]). *Élémens de la grammaire françoise.* Colas.
Restaut P. (1745 [1732]). *Abrégé des principes de la grammaire françoise.* Chez l'Auteur.
Wailly N.-F. de. (1759). *Abrégé de la grammaire françoise.* Barbou.

Les erreurs interférentielles et le portfolio linguistique du locuteur, deux variables corrélées par le rho de Spearman

Lucia Ráčková

Matej Bel University in Banská Bystrica – lucia.rackova@umb.sk

Abstract
In Slovak produced by Native Francophone speakers, it is possible to observe the phenomenon of interference, especially in its verbal system. Using correlations analysis, also called Spearman's ρ, we try to answer research questions linked to authentic statements including linguistic features of other languages. This method which consists of looking for a correlation coefficient between two variables, is the most suitable for evaluating the closed questions of the questionnary-test.

Keywords: Verbal Interference, Francophone users of Slovak, Language Portfolio, Spearman's ρ

Résumé
Dans le slovaque produit par les locuteurs francophones natifs, il est possible d'observer le phénomène de l'interférence, notamment dans son système verbal. En nous servant de l'analyse de corrélations, appelée également le ρ de Spearman, nous cherchons à répondre aux questions de recherche liées aux énoncés authentiques comprenant les traits linguistiques d'autres langues. Cette méthode, qui consiste à chercher un coefficient de corrélation entre deux variables, est la plus convenable pour évaluer les questions fermées du questionnaire-test.

Mots-clés: interférence verbale, locuteurs francophones de slovaque, portfolio linguistique, le ρ de Spearman

1. Introduction

Cette communication prend comme point de départ des énoncés authentiques produits par des sujets francophones en cours d'apprentissage du slovaque. Ces textes ont été recueillis par la méthode du questionnaire dont nous nous sommes servie dans notre thèse qui se concentre particulièrement sur le système verbal du slovaque.

En faisant usage de la méthode statistique du rho de Spearman, outil statistique d'une mesure de dépendance statistique non paramétrique entre deux variables, convenable pour le type des variables présentes dans notre enquête par

questionnaire à questions fermées et partiellement pour les questions ouvertes, nous cherchons à répondre aux quatre questions principales pour lesquelles les corrélations de Spearman devraient être significatives. Premièrement, la connaissance d'autres langues slaves auparavant étudiées par les sujets plurilingues exerce-t-elle une influence sur la qualité des communications en slovaque ? Nous émettons l'hypothèse d'une influence positive de la maîtrise d'autres langues slaves sur les locuteurs francophones de slovaque, particulièrement en ce qui concerne les formes de flexion typiques pour la branche slave des langues. Deuxièmement, nous nous interrogeons sur l'impact d'autres langues présentes dans le bagage linguistique des locuteurs francophones du slovaque. Il est à supposer que chaque langue étudiée auparavant facilite l'acquisition d'une nouvelle langue.
La dernière question qui nous intéresse est de savoir si les locuteurs francophones en cours d'apprentissage du slovaque savent évaluer leur niveau de slovaque à l'écrit et à l'oral. Nous partons de l'idée que leur estimation du niveau de la langue apprise n'est pas objective.

2. Cadre théorique

Dans le cadre théorique, nous commencerons par une brève présentation de l'outil statistique du ρ de Spearman,utilisable pour la recherche en science du langage. Ensuite, nous développerons les questions du portfolio linguistique et du transfert négatif des connaissances d'une langue vers l'autre.
Le coefficient de Spearman, {\displaystyle \rho }une méthode employée en statistique, sert à mettre en liaison deux variables bien distinctes. Elle permettra au chercheur de parvenir à des corrélations qui pourront ou non être significatives. Sa valeur est comprise entre *-1* et *1*. Un *coefficient* de corrélation *de Spearman* de *1 ou de -1* est obtenu quand *les* deux variables comparées ont une relation parfaitement monotone. Une valeur nulle indique que *les* deux classements n'ont rien à voir l'un par rapport à l'autre.
Il est évident que la langue maternelle a un impact fort sur l'acquisition de la langue seconde de l'étudiant et bien sûr sur d'autres langues apprises. « En effet, chez le locuteur polyglotte, différents systèmes linguistiques entrent en interaction, en particulier au niveau morphosyntaxique. » (Ráčková et Schmitt, 2019 : 9). La langue maternelle et les langues étrangères présentes dans le portfolio linguistique se projettent, par conséquent, sur les structures de la langue slovaque. Ce transfert peut être positif ou négatif. De plus, le slovaque est une langue à flexion externe, donc une langue très complexe (Feuillet, 2006 : 23).

3. Analyse et résultats

Nous avons appliqué la méthode de Spearman, par une analyse de corrélation effectuée par le programme IBM SPSS version 19, sur les résultats obtenus par

questionnaire-test proposé aux locuteurs. Comme les francophones en train d'acquisition du slovaque ne sont pas nombreux, nous avons recueilli 69 questionnaires remplis. Nous avons codé les cases cochées obtenues des données nominales (la connaissance d'autres langues slaves, l'impact d'autres langues étudiées auparavant, niveau en langue estimée), mises en relation avec les résultats du test de connaissance, en données numériques. Ce qui est évident est que, par la méthode du ρ de Spearman, il est beaucoup plus facile d'évaluer les questions fermées que les questions ouvertes à réponses libres. « Mais les questions ouvertes constituent un prolongement indispensable des questionnaires lorsque les enquêtes vont au-delà d'une simple recherche de suffrages, lorsqu'il s'agit d'explorer et d'approfondir un sujet complexe ou mal connu » (Lebart et Salem, 1994 : 23). Le regroupement de réponses reçues dans notre recherche se fait par le classement des réponses erronées dans les catégories des transferts sémantique et grammatical. Ce qui est important pour notre recherche, c'est que les réponses aux questions ne soient pas trop variées pour que les conclusions ne prennent pas un caractère subjectif et soient, de ce fait, difficilement généralisables.

Avant de procéder à l'analyse, nous avons ramassé les données de nature sociolinguistique par le questionnaire-test disponible en ligne. Celui-ci comportait deux parties principales : des questions portant sur les données de caractère social des locuteurs et un test de slovaque langue étrangère.

Pour juger de l'impact des langues apprises antérieurement par les locuteurs sur la nouvelle langue en cours d'apprentissage, nous avons mis en corrélation deux variables : les données sur la représentation des langues étrangères dans le portfolio des apprenants et le niveau de slovaque estimé par les apprenants eux-mêmes et dont il nous ont rendu compte dans le questionnaire. Ensuite, nous avons cherché les corrélations entre les variables de notre recherche.

Nous présenterons de manière successive les corrélations significatives que notre recherche a révélées. Premièrement, il s'agit de la corrélation entre les connaissances antérieures dans les langues slaves avec les résultats obtenus des locuteurs dans le test écrit. La première concernait la corrélation entre la maîtrise d'une autre langue slave par les locuteurs et les résutats du test écrit. Le coefficient de -0,2 nous permettait de constater une dépendance indirecte entre les deux éléments, ce qui signifie que le nombre d'erreurs commises par les locuteurs francophones de slovaque qui possèdent une autre langue slave (le polonais, le russe, le serbo-croate, le tchèque ou l'ukrainien) dans leur portfolio linguistique est moins important qu'auprès des locuteurs pour qui le slovaque représente la première langue slave étudiée.

La deuxième question concernait l'influence des langues étrangères antérieurement étudiées sur le slovaque langue étrangère. Pour rendre notre recherche objective, nous nous sommes centrée sur les langues étrangères les plus représentées dans les portfolios : l'anglais, l'allemand et l'espagnol. La corrélation

entre la maîtrise de l'anglais et les erreurs grammaticales et sémantiques constatées ont donné un coefficient de -0,254. Cela veut dire que les locuteurs francophones qui parlent anglais ont un niveau plus bas que ceux qui ne parlent pas cette langue, même s'il ne s'agit que d'un petit sous-groupe de notre échantillon représentatif.

En nous servant de la méthode du rho de Spearman, nous sommes tombée sur des associations intéressantes, surtout en comparant les langues allemande et espagnole dans le portfolio des apprenants. Les locuteurs qui maîtrisent l'allemand ont un niveau plus élevé en langue slovaque que leurs collègues qui n'ont jamais appris cette langue germanique avec un coefficient de -0,268. Par contre, les locuteurs ayant appris auparavant l'espagnol, ont obtenu moins de bonnes réponses que le reste de l'échantillon. Le coefficient de Spearman de -0,335 nous indique que les individus qui parlent espagnol commettent plus d'erreurs en langue slovaque que les francophones qui ne parlent pas cette langue.

La troisième question qui nous intéressait était la langue de communication du couple franco-slovaque. Nous avons découvert que, si la langue de communication du couple est l'anglais, le niveau de slovaque des francophones concernés est plus faible (le coefficient de Spearman est égal à -0,55). La dernière question concernait le niveau de slovaque à l'écrit et à l'oral des locuteurs français. Il s'agissait de leur propre estimation de maîtrise du slovaque. Les résultats de l'analyse de corrélations démontrent que les locuteurs ont une image réelle de leur niveau en slovaque. Pour la production orale, le programme utilisé a montré une forte dépendance directe (le coefficient de Spearman est égal à -0, 701). Nous avons fait la même constatation pour la production écrite (le coefficient de Spearman est égal à -0, 679). Il est à noter que plus les corrélations se rapprochent de 0, moins significatives elles sont.

4. Conclusion

L'anglais, maîtrisé par la majorité des locuteurs, a marqué de manière négative la qualité des énoncés slovaques, de même que l'espagnol. Par contre, les locuteurs francophones de slovaque qui parlent l'allemand maîtrisent mieux le slovaque. Nous expliquons ce phénomène intéressant par la typologie de l'allemand qui, bien que langue germanique et analytique, comprend également une flexion assez riche, ce qui facilite l'apprentissage ultérieur des langues flexionnelles. Ces résultats mériteraient à être développés davantage par la mise en corrélation d'autres variables.

Bibliographie

Feuillet J. (2006). *Introduction à la typologie linguistique*. Honoré Champion.

Lebart L. et Salem A. (1994). *Statistique textuelle.* Dunod.

Ráčková L. et Schmitt F. (2019). *Les interférences linguistiques du français sur le slovaque. L'exemple du systéme verbal.* L'Harmattan.

Retrouver l'inventeur-auteur: la levée d'homonymies d'autorat entre les brevets et les publications scientifiques

David Reymond[1], Manuel Durand-Barthez[2],
Heman Khouilla[3], Sandrine Wolff[4]

[1]Université de Toulon, Aix Marseille Univ, IMSIC -
dreymond@univ-tln.fr
[2]URFIST Paris - manuel.durand-barthez@orange.fr
[3]Université de Toulon, LÉAD –
heman-khouilla-koumba@etud.univ-tln.fr
[4]Université de Strasbourg, Université de Lorraine, CNRS, BETA –
wolff@unistra.fr

Abstract
Patents and scientific papers provide an essential source for measuring science and technology output, to be used as a basis for the most varied scientometric analyzes. Authors' and inventors' names are the key identifiers to carry out these analyses, which however, run up against the issue of disambiguation. By extension identifying inventors who are also academic authors is a non-trivial challenge. We propose a method using the International Patent Classification (IPC) and the IPCCAT API to assess the degree of similarity of patents and papers abstracts of a given inventor, in order to match both types of documents. The method is developed and manually qualified based on three corpora of patents extracted from the international EPO database Espacenet. Among a set of 4679 patents and 7720 inventors, we obtain 2501 authors. The proposed algorithm solves the general problem of disambiguation with an error rate lower than 5%.

Keywords: disambiguation; author; inventor; homonymy; patent; publication

Résumé
Brevets d'invention et articles scientifiques représentent une source essentielle pour mesurer la production scientifique et technologique, servant de base aux analyses scientométriques les plus variées. Les noms d'auteurs et d'inventeurs constituent un identifiant pivot pour réaliser ces analyses, mais ils nécessitent d'être désambiguïsés. Par extension l'identification des inventeurs qui sont également des auteurs académiques est un problème non trivial. Nous proposons une méthode utilisant la classification internationale des brevets et l'API IPCCAT pour évaluer le degré de similarité entre le résumé du brevet d'un inventeur donné et ceux des articles de personnes homonymes,

afin d'apparier des deux types de documents. La méthode est développée et qualifiée manuellement sur trois corpus de brevets issus de la base internationale Espacenet de l'OEB. Sur un jeu de 4679 brevets et 7720 inventeurs, nous obtenons un résultat de 2501 auteurs. L'algorithme proposé résout le problème général de la levée d'homonymie avec un taux d'erreur inférieur à 5%.

Mots clés : désambiguïsation ; auteurs ; inventeur ; homonymie ; brevets ; publications

Remerciements : Ce travail a bénéficié d'un financement du GIS Réseau URFIST dans le cadre du projet Brevets : http://gis-reseau-urfist.fr/brevets/, ainsi que les pôles Information, Prévention, Numérique et Santé et Pôle Mer Environnement et Développement Durable (MEDD) de l'Université de Toulon.

1. Introduction

Les brevets d'invention comme les articles scientifiques, en tant que produits identifiables et mesurables des activités de recherche et développement, constituent la source informationnelle la plus répandue pour analyser la recherche et l'innovation, en particulier les interactions entre la Science et l'Industrie[1]. Les noms d'auteurs ou d'inventeurs constituent un identifiant pivot pour la réalisation d'analyses sur ces questions, mais ils comportent des problèmes de désambiguïsation notamment par rapport à l'homonymie, c'est-à-dire une même forme lexicale pour deux personnes différentes[2], étant entendu que le point de départ est constitué de bases de données accessibles librement et, en général, non nettoyées (Laender et al., 2008; Kim et al., 2014; Hussain & Asghar, 2017).

En utilisant la classification internationale des brevets (CIB[3]) et l'API IPCCAT comme instruments de mesure de la similarité des productions associées à un nom, nous proposons une nouvelle méthode de levée d'homonymie d'autorat, que nous expérimentons sur trois corpus de brevets extraits de la base internationale Espacenet de l'Office Européen des Brevets (OEB). La qualification des résultats de l'algorithme a été effectuée manuellement via la méthode d'échantillonnage statistique. Sur un ensemble de 4679 brevets correspondant à 7720 inventeurs, 2501 parmi eux sont aussi des auteurs. L'algorithme proposé résout le problème général de la levée d'homonymie avec un taux de réussite très satisfaisant (<5% d'erreur).

[1] Cf. Foray& Lissoni (2010) pour une synthèse générale sur les relations universités-entreprises.

[2] La désambiguïsation des noms de déposants (Dassa et al. 2014) n'est pas abordée ici, car pour nos expériences elle est réalisée à l'aide d'un dictionnaire fourni par l'OEB, et une table de correspondance, que nous avons développée en extension de la ressource fournie par l'OEB (Office 2019).

[3] Ou *International Patent Classification* (IPC) en anglais.

Au sein des trois corpus, notre objectif est d'identifier, parmi les noms d'inventeurs, ceux qui sont aussi des auteurs de publications, fort probablement issus du monde académique. La méthodologie est construite en trois phases. La première regroupe au niveau lexical les inventeurs. La seconde collecte dans une base complémentaire les données des publications académiques associées à chaque nom d'inventeur. Enfin, la dernière étape consiste à rapprocher le résumé des publications associées à ce nom (qu'il s'agisse d'un homonyme ou d'un potentiel inventeur-auteur) avec le descriptif du contenu du brevet, ce qui doit permettre *in fine* de déterminer si l'inventeur est bien l'auteur des publications.

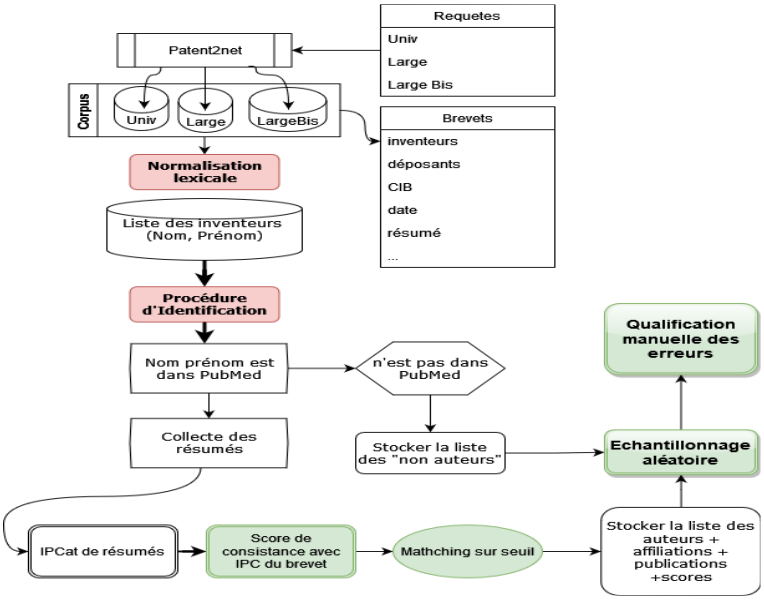
2. Construction des jeux de données

Pour parvenir à notre objectif, nous avons construit à l'aide du logiciel *Patent2Net* (Reymond 2021) trois corpus de demandes de dépôts de brevets, issus de la base mondiale des brevets worldwide.espacenet.com hébergée par l'OEB. Les requêtes se limitent aux demandes de priorité française (PR=FR) et ciblent un domaine technologique particulier et une période[4].

- **Univ** : Vise des brevets détenus par des organismes de recherche publique ("univ* ou laboratoire ou institut" dans le nom du déposant) sur la période 2014-2016. La requête est : "(pa=univ* or pa=institut or pa=laboratoire) AND (ic=A61) AND (pd within "2014, 2016") AND pr=FR". Le domaine technologique choisi (A61) cible, dans la grande catégorie des nécessités courantes de la vie, le domaine médical ou vétérinaire et l'hygiène. On y verra majoritairement des déposants universitaires, acteurs de la valorisation du domaine.
- **Large** : Même période, mais ne cible pas des déposants spécifiques et vise deux domaines technologiques **différents** (A63 - sports, jeux…- et A62 – sauvetage ; lutte contre les incendies), sans imposer de contrainte sur le nom de déposant *a priori.* La requête est : "*(ic=A63 OR IC=A62) AND (pd within "2014, 2016") AND pr=FR*".
- **LargeBis** : Même domaine que **Univ**, mais sur une **autre période** proche (2013). La requête est : "ic=A61 AND pd=2013 AND pr=FR". Se distingue de **Univ** par la levée de la contrainte sur le nom de déposant.
-

[4] Les critères de sélection des corpus visent à lever les biais technologiques et temporels tout en appréciant l'efficacité de la méthode.

3. Procédure d'extraction



La Figure 1 présente le fil des traitements réalisés sur les différents corpus est détaillé ci-après :

1 Une pré-normalisation s'appuie sur la bibliothèque python *fuzzy-wuzzy*[5] avec une distance de Levenshtein insensible à la position des mots pour rapprocher des chaînes lexicales (inversions noms-prénoms, coquilles). La chaîne la plus longue est privilégiée. Cette étape construit la liste des inventeurs.

2 Procédure d'identification de chaque inventeur en deux temps :
   a Interrogation de l'API PubMed pour récupérer l'ensemble des publications associées au dit inventeur. L'outil de désambiguïsation de PubMed (Torvik *et al.* 2005 ; Lerchenmueller et Sorenson 2016) permet de chercher les variations d'une forme lexicale.
   b Chaque production identifiée est rattachée à l'inventeur. Cette étape produit la liste des auteurs homonymes aux inventeurs.

3 L'API IPCCAT (Fiévet et Guyot 2018a, 2018b) est ensuite utilisée ensuite de manière originale[6] pour rapprocher le résumé de chaque publication scientifique d'auteur homonyme au résumé du brevet

[5] https://github.com/seatgeek/thefuzz

[6] L'API IPCCAT est utilisée en général pour identifier le code CIB de textes décrivant des brevets d'invention, ce qu'elle fait avec un très bon niveau de fiabilité d'après les auteurs.

correspondant, en indiquant pour chaque texte les classes technologiques de la CIB identifiées. Les scores de pertinence des CIB retournés par l'API sont alors utilisés pour associer les auteurs :

a Si le score est >800[7], et si l'un des codes CIB (5 digits) attribués au brevet de l'inventeur figure parmi ceux associés par IPCCAT aux publications scientifiques, alors l'inventeur est considéré comme l'auteur des publications.

b Dans ce cas, un appel à l'API de géoréférencement avec l'adresse fournie par Pubmed (pas toujours complète) vient compléter l'information auteur.

4 Une opération manuelle de vérification sur le Web (LinkedIn, Research Gate, Google Scholar…) est effectuée par la suite sur 10% des auteurs prélevés aléatoirement, afin de contrôler l'autorat.

Implanté en open source (https://github.com/Patent2net/P2N-v3/tree/Prg7), le procédé est flexible : la procédure d'identification (étape 2, figure 1 peut-être notamment adaptée à la collecte de publications issues de bases bibliographiques pluridisciplinaires (le Web of Science, Scopus, etc.).

4. Mode de qualification des données et résultats

L'échantillon aléatoire total sur les 3 groupes comprend 390 auteurs-inventeurs candidats parmi les 2501 traités par IPCCAT. Le risque d'homonymie est évalué avec une marge d'erreur < 5% (Thelwall 2009, 29-31). La vérification manuelle sur les sites universitaires et les réseaux sociaux corrobore les effets de la désambiguïsation : 96% sur les 3 corpus. 17 cas font exception (parmi lesquels 15 font doute et moins de 1% relève d'une erreur certaine - soit 0,0005 %)
Le Tableau 1 regroupe ces éléments de qualification :

Tableau 1. Résultats de qualification sur les échantillons aléatoires des candidats auteurs-inventeurs

Corpus	Univ	Large	LargeBis	Total
Inventeurs	2122	1088	4510	7720
Auteurs-inventeurs candidats (Correspondance IPCCAT seuil > 800)	1209	27	1265	2501
Echantillon	137	8	244	390
Candidats vérifiés	133 (97 %)	8	231 (95 %)	373 (96 %)
Doute	3	0	12	15
Erreur	1	0	1	2 (0.0005 %)

[7] Les auteurs d'IPCCAT suggèrent de veiller à un score > 1200 pour garantir la fiabilité des réponses dans le cas des textes de brevets. Notre utilisation détournée d'IPCCAT sur des productions académiques nous conduit à être plus souples quant au seuil utilisé : d'une part le style rédactionnel des articles scientifiques diffère de celui d'un brevet, d'autre part un seuil plus bas permet une tolérance plus vaste quant à la connexité des travaux.

Le Tableau 2 ci-après synthétise les résultats du processus de levée d'homonymie sur les trois corpus. Dans le domaine médical, avec la base PubMed, la correspondance IPCCAT sur **Univ** et **LargeBis** (càd sans imposer la présence d'un déposant académique) montre que **dans les deux cas** plus de 70% des inventeurs homonymes sont candidats auteurs-inventeurs (càd probablement académiques).

Tableau 2. Inférence des proportions auteurs-inventeurs par levée d'homonymie et ratio d'autorat

Corpus	**Univ**	**Large**	**LargeBis**
Nombre de brevets	683	874	3122
Nombre d'inventeurs normalisés uniques	2122	1088	4510
Nombre d'auteurs homonymes (avant désambiguïsation)	1591	182	1765
Auteurs-inventeurs candidats (Correspondance IPCCAT seuil > 800)	1209	27	1265
Proportion d'auteurs-inventeurs matchée (après traitement IPCCAT) sur les auteurs homonymes	74.78	14.89	71.67

5. Conclusion

Ce procédé de désambiguïsation, assimilable à un rapprochement sémantique (Block et al. 2021), établit l'autorat effectif dans plus de 95% des cas, le contrôle manuel sur échantillon aléatoire des auteurs-inventeurs candidats ne révélant que 5% d'erreur. Le groupe **Univ** (déposants académiques) traité avec IPCCAT conserve 74,78% d'homonymes pertinents. Le score similaire (71.67 %) dans le jeu étendu sur les demandes de brevets de l'année précédente (**LargeBis**) montre la stabilité du procédé et sa fiabilité. La focalisation sur une spécialité scientifique est optimisée par le choix d'une base de données source réellement adéquate (ici : PubMed). La méthode expérimentée ici, dans un contexte de généralisation de la science ouverte, pourrait s'étendre à tout autre domaine scientifique lié à l'industrie.

Le rapprochement des demandes de brevet et des publications par ce procédé met en évidence le « prior art », non seulement pour vérifier la nouveauté de ces demandes mais bien plus pour améliorer la visibilité de la composante « recherche » dans l'avènement de technologies nouvelles (Lissoni et Montobbio 2015) sans recourir par défaut aux seules citations bibliographiques (Verbeek et al. 2002 ; Acosta et Coronado 2003 ; Glänzel et Meyer 2003 ; Ahmadpoor et Jones 2017).

Des travaux antérieurs (Reymond, Galliano et Quoniam 2019) nous ont permis d'esquisser la pertinence de la classification internationale des brevets comme pivot de l'interdisciplinarité sur la base des thèses en ligne. Cette deuxième approche, bien plus concluante malgré les limites d'erreurs marginales, laisse en

outre envisager des utilisations et extensions en scientométrie en tant qu'instrument fiable de décloisonnement des sciences, de l'industrie et de la société.

Bibliographie

Acosta M. and Coronado D. (2003). Science–technology flows in Spanish regions: an analysis of scientific citations in patents ». *Research policy* 32 (10). Elsevier : 1783-1803.

Ahmadpoor M. and Jones B.F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science* 357 (6351). American Association for the Advancement of Science : 583-587.

Block C., Wustmans M., Laibach N. and Bröring S. (2021). Semantic Bridging of Patents and Scientific Publications – The Case of an Emerging Sustainability-Oriented Technology. *Technological Forecasting and Social Change* 167 (juin): 120689. https://doi.org/10.1016/j.techfore.2021.120689.

Dassa M., Deniau A., Durand-Barthez M., Girard F., Pothier N. and Sene. A. (2014). Le millefeuille des affiliations françaises dans les publications scientifiques. *Documentaliste-Sciences de l'Information* Vol. 51 (4) : 12. https://doi.org/10.3917/docsi.514.0012.

Fiévet P. and Guyot F. (2018a). Automatic Categorization of Patent Documents in the International Patent Classification (IPCCAT). In Nice : *WIPO IC-SDV*. https://haxel.com/ii-sdv/2018/Programme/monday-23-april-2018.

Fiévet P. and Guyot F. (2018b). Text categorization in the International Patent Classification at group level.In *WIPO IC-SDV*.
https://www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/Documentation/presentations/20180423_IPCCAT_ICSDV_2018.pdf.

Foray D. and Lissoni F. (2010). University research and public–private interaction. Dans *Handbook of the economics of innovation*, 1 : 275-314. Elsevier.

Glänzel W. and Meyer M. (2003). Patents cited in the scientific literature: an exploratory study of'reverse' citation relations. *Scientometrics* 58 (2) : 415-428.

Hussain I., and Asghar S. (2017). A survey of author name disambiguation techniques: 2010-2016. *Knowledge Eng. Review*, 32, e22.

Kim J., Diesner J., Kim H., Aleyasen A. and Kim H.-M. (2014). Why name ambiguity resolution matters for scholarly big data research. *2014 IEEE International Conference on Big Data (Big Data), 1-6.* https://doi.org/10.1109/BigData.2014.7004345

Laender A.H., Gonçalves M.A., Cota R.G., Ferreira A.A., Santos R.L. and Silva A.J. (2008). Keeping a digital library clean: New solutions to old problems ». Proceedings of the eighth ACM symposium on Document engineering, 257-262.

Lerchenmueller M.J. and Sorenson O. (2016). Author disambiguation in PubMed: Evidence on the precision and recall of author-ity among NIH-funded scientists. *PLoS One* 11 (7). Public Library of Science San Francisco, CA USA : e0158731.

Lissoni F. and Montobbio F. (2015). Guest authors or ghost inventors? Inventorship and authorship attribution in academic science. *Evaluation Review* 39 (1). SAGE Publications Sage CA: Los Angeles, CA : 19-45.

Office, European Patent (2021). Espacenet: Patent Database with over 130 Million patent

Documents, septembre. https://www.epo.org/searching-for-patents/technical/espacenet.html#tab-1.
Reymond D. (2021). *Patent2Net version 3* (version 3). Docker. Python.
Reymond D., GallianoC. and Quoniam L.. (2019). La CIB comme pivot de classement interdisciplinaire. In *Faire dialoguer les disciplines via l'indexation des connaissances : la recherche interdisciplinaire en débats*. Montpellier, France. http://numerev.com/programme-colloque-numerev.html.
Thelwall M. **(**2009). *Introduction to webometrics. Quantitative web research for the social sciences*. Synthesis Lectures on Information Concepts Retreival and Services. Morgan & Claypool Publishers.
Torvik, V.I., Weeber M., Swanson D.R. and Smalheiser N.R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology* 56 (2). Wiley Online Library : 140-158.
Verbeek A., Debackere K., Luwel M., Andries P., Zimmermann E. and Deleus F. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics* 54 (3). Springer : 399-420.

La stabilité du cadre discursif du Ministère des Affaires Etrangères à l'épreuve de la textométrie

Annabel Richeton, Virginie Lethier, Margareta Kastberg Sjöblom

Université de Bourgogne Franche-Comté – annabel.richeton@edu.univ-fcomte.fr virginie.lethier@univ-fcomte.fr margareta.kastberg@univ-fcomte.fr

Abstract
The communication of the Ministry of Foreign Affairs is submitted to standards of homogeneity and coherence albeit carried by multiple voices. The ministry speeches are also at the interface of institutional and political discourses. We investigate factors of variation involved in this type of discourse: the place of the speaker, their political affiliation, or his status. To test these hypotheses, we propose a textometric exploration of a large corpus (4,191 texts) produced between 1997 and 2019 (5 presidential mandates and 36 speakers holding three different functions). Correspondence analyses point out these variations, by analysing the distribution of parts of speech by mandate and by speakers.

Keywords: Discourse analysis; Institutional discourse; Ministry of Foreign Affairs; Metadata

Résumé
La communication du Ministère des Affaires étrangères est soumise à des normes d'homogénéité et de cohérence, mais portée par des voix multiples. Son discours se trouve à l'interface des discours institutionnel et politique. Nous nous intéressons aux facteurs de variations intervenant dans ce type de discours : la place du locuteur, son affiliation politique ou son statut. Pour tester ces hypothèses, nous proposons l'exploration textométrique d'un vaste corpus de 4.191 textes produits entre 1997 et 2019 (cinq mandats présidentiels et 36 locuteurs occupant 3 fonctions différentes). Des analyses factorielles des correspondances mettent au jour ces variations, en interrogeant la ventilation des parties du discours par mandat et par locuteur.

Mots clés : Analyse du discours ; Discours institutionnel ; Ministère des Affaires étrangères ; Métadonnées

1. Introduction

Dans la continuité des travaux sur le discours institutionnel (Gobin et Deroubaix, 2010, entre autres) et sur le discours politique (Mayaffre, 2004), nous proposons une exploration outillée du discours du Ministère des Affaires étrangères (MAE) français. En effet, c'est une position d'interface entre discours politique et discours institutionnel qu'occupe ce cas d'étude : produit par un ministère régalien, il relève du discours politique : le locuteur doit faire entendre la politique du gouvernement auquel il appartient, parler ou agir « au nom d'une entité qui nous dépasse et qui nous a délégué – provisoirement – ce pouvoir de parler en son nom » (Charaudeau, 2005 : 53). Il est aussi un discours institutionnel : en dépit de la diversité de ses représentants, l'objectif de ce type de discours est de faire entendre une voix unique et univoque afin d'être lisible, audible, crédible, et légitime (Oger et Ollivier-Yaniv, 2006).

Dès lors, on peut s'interroger sur les marques de l'individualité du locuteur dans un discours par essence homogène et régi par des normes strictes. Autrement dit, peut-on observer dans le discours du MAE des variations liées au locuteur en tant qu'individu ? Pour répondre à cette question, nous avons construit un corpus intégrant toutes les déclarations disponibles en ligne[1] entre juin 1997 – date du début de la cohabitation Chriac/Jospin – et décembre 2019 afin d'observer une partie du mandat de Macron. En effet, le corpus doit pouvoir être partionné en mandats, qui correspondent à une période mais surtout à une tendance politique induite par le président et le chef du Gouvernement. Ce corpus a été soumis à une analyse textométrique (Heiden, 2010) en vue de dégager de grandes tendances et d'observer des oppositions entre les différentes métadonnées qui enrichissent notre corpus.

2. Contexte de la recherche

2.1. Le MAE et sa communication

Le MAE a pour mission de « représenter, défendre et promouvoir les intérêts de la France et des ressortissants français dans tous les domaines auprès des pays étrangers et au sein des organisations internationales »[2]. Si l'un des principaux versants de son activité communicative, tourné vers ses homologues étrangers, n'est pas accessible au chercheur, il est possible d'observer le discours du MAE à travers sa communication externe : les prises de paroles publiques de ses ministres, ministres délégués et secrétaires d'Etat. Ces prises de parole peuvent s'adresser à des publics-cibles (ambassadeurs, professionnels spécialistes,

[1] La DILA diffuse les transcriptions des prises de parole des acteurs de la vie publique (www.vie-publique.fr)

[2] https://www.diplomatie.gouv.fr/fr/le-ministere-et-son-reseau/missions-organisation/

journalistes ou acteurs politiques), mais sont également produites à destination du grand public : l'objectif est de faire connaître et de défendre, auprès plus grand nombre, les actions du ministère. Bien que prononcés par la voix d'un représentant unique, ces discours résultent d'une chaîne d'écriture collective, de tours d'écriture entre les conseillers·ères spécialisé·e·s, les directions concernées, la plume du ministre et le ministre. Il s'agit de coproductions dans lesquelles plusieurs voix se succèdent pour créer un document cohérent et homogène, garants de la crédibilité de la parole des locuteurs.

2.2. Hypothèses et questions de recherche

Le discours du MAE étant un discours coproduit mais destiné à être porté par une voix, à circuler dans différentes sphères en étant compris par chacun, et devant être suffisamment décontextualisé pour être reproduit en toutes circonstances (Oger et Ollivier-Yaniv, 2006), peut-on y repérer une empreinte individuelle du locuteur ? Leur identité socio-professionnelle, leur affiliation politique ou encore leur genre peuvent-ils jouer un rôle dans la construction de leurs discours ? Nous nous interrogeons également sur la présence de traces d'une évolution des fonctions de l'institution dans les discours qu'elle produit, selon les mandats présidentiels, les gouvernements, les représentants ou les périodes. Une autre hypothèse consiste donc à considérer le poids du contexte politique des productions discursives du ministère.

2.3. Corpus

Notre corpus contient ainsi 4.191 textes – soit 6.588.488 occurrences – prononcées par 36 locuteurs. Nous nous concentrons sur les prises de parole monologales intitulées « Déclaration » lors de leur mise en ligne. La déclaration est l'« action de déclarer, de porter à la connaissance du public » et par extension un « discours par lequel on fait publiquement une communication »[3]. Dans une perspective politique, elle « exprim[e] la prise de position, la décision officielle d'un chef ou d'un groupe vis-à-vis d'un autre groupe »[4]. Ce corpus de travail est donc homogène en genre et les prises de paroles y sont ainsi réparties :

Mandat	Parti politique du président	Parti politique du gouvernement	Parti politique des ministres	Nombre d'occurrences
Chirac1 (1997-2002)	RPR	PS	PS	1 366 132
Chirac1 (2002-2007)	RPR-UMP	RPR-UMP	RPR-UMP	1 247 987
Sarkozy (2007-2012)	UMP	UMP	PS-UMP	1 405 692
Hollande (2012-2017)	PS	PS	PS	1 766 720
Macron (2017-2019)	LREM	LR	PS	769 099

[3] Larousse en ligne. URL: https://www.larousse.fr/dictionnaires/francais/d%C3%A9claration/22224

[4] CNRTL en ligne. URL: https://www.cnrtl.fr/definition/déclaration

3. La construction du discours dans l'échiquier politique : le cas des déclarations.

Pour observer une évolution dans la construction du discours du MAE entre 1997 et 2019, le corpus est découpé en 5 parties (cf. tableau ci-dessus), intégrant les déclarations de locuteurs ayant trois statuts différents : ministres, ministres délégués et secrétaires d'Etat. Le choix de l'unité d'indexation et de segmentation des catégories grammaticales, issu d'une annotation automatique par Treetagger[5], témoigne de notre volonté d'éviter une entrée lexicale plus sensible aux phénomènes conjoncturels. Ici, les différents types de pronoms (relatifs, possessifs, personnels, démonstratifs et indéfinis) ont été réunis, de même que les déterminants (articles et possessifs). Les noms communs et propres ont été réunis. Sur la base de cette matrice lexicale, une AFC a été générée, résumant près de 94 % de l'information du tableau :

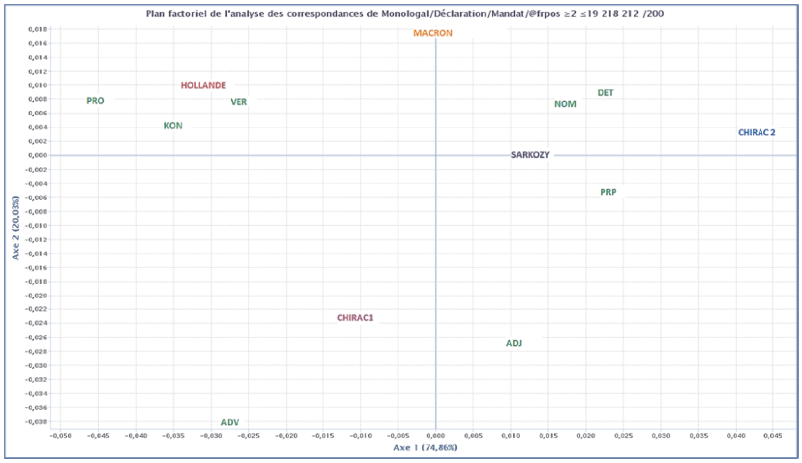


Figure 1. Plan factoriel de l'analyse des correspondances de Monologal/Déclaration/Mandat/@frpos > 19 218 212/200

Sur l'axe 1 (74,86 % de l'information), nous observons une opposition nette entre le pôle nominal (noms, déterminants, adjectifs et prépositions) et le pôle verbal (pronoms, verbes, adverbes et conjonctions), opposition fréquente dans les corpus clos et amplement documentée par Brunet (1981). Le pôle verbal (partie ouest) attire les ministres de Hollande et de Chirac lors de son premier mandat ; le pôle nominal, à l'est, attire les ministres de Sarkozy et de Chirac (second mandat). La configuration des points colonnes n'atteste pas d'un poids de la diachronie comme nos hypothèses le formulaient, mais rapprochent les

[5] Nécessairement imparfaite (Haber, 2004)

mandats en fonction de la politique d'appartenance de leur Gouvernement. En effet, les représentants du ministère entre 1997 et 2002 (Chirac1) sont des ministres socialistes, tout comme les ministres sous Hollande.
Nous pouvons donc observer que les ministres socialistes semblent privilégier des constructions verbales tournées vers l'action, tandis que les ministres de droites sont polarisés autour d'un discours plus nominal et donc tourné vers l'abstraction et la théorisation (Duchastel, 1993).
L'axe 2 (20 % de l'information) montre que les ministres du mandat Chirac1 suremploient les adverbes. Ce goût pour l'adverbe des représentants du MAE n'est pas sans faire écho aux travaux de Mayaffre (2004) pointant la construction du discours di président Chirac[6].

4. Du poids des statuts

Pour tester l'hypothèse du poids de l'identité des locuteurs (genre, statut, affiliation politique, etc.), nous avons partitionné notre corpus par locuteur. Nous avons ainsi construit une matrice lexicale croisant en lignes les mêmes parties du discours (noms, adjectifs, déterminants, prépositions, pronoms, verbes, adverbes et conjonctions) et en colonnes les 39 locuteurs du corpus, et générer une AFC portant près de 92,5 % de l'information de cette matrice.

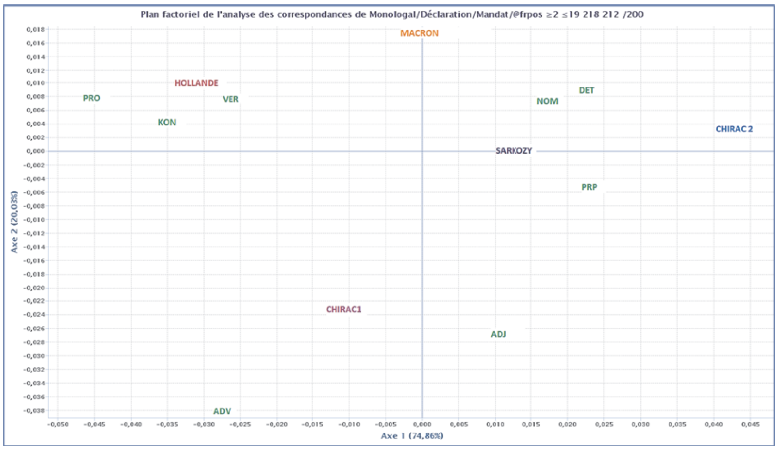


Figure 2. Plan factoriel de l'analyse des correspondances de Monologal/Déclaration/Locuteur/@frpos> 19 218 212 /200

Sur l'axe 1, l'opposition la plus nette concerne ici le statut de chaque locuteur. En effet, 5 des 9 ministres (en majuscules sur le graphique) se trouvent dans la

[6] Chirac tend à suremployer l'adverbe « naturellement ». Nous retrouvons cet adverbe parmi les plus employés par les représentants du MAE sous son premier mandat. « Naturellement » sous-entend, dans leur discours, l'évidence et le caractère essentiel – voire logique – de l'objet de la phrase.

partie ouest. De manière plus franche encore, seuls 4 des 25 ministres délégués ou secrétaires d'Etat se trouvent dans la partie ouest, quand 20 se trouvent à l'est (et une sur l'axe des ordonnées).

L'opposition en fonction des statuts est tout particulièrement nette pour le mandat Hollande (en rose sur le graphique) : les 2 ministres, Ayrault et Fabius, sont à gauche, et les 8 ministres délégués (Benguigui, Cazeneuve, Girardin, Désir, Canfin, Repentin, Vallini et Le Guen) droite du graphique. En termes discursifs, cette opposition traduit deux modes d'énonciation distincts : le discours des ministres privilégie une énonciation prise en charge, tandis ceux des ministres délégués et secrétaires d'Etat favorisent une énonciation plus factuelle et objective.

Dans certains extraits, on observera que le discours du secrétaire d'Etat Harlem Désir donne le primat aux chiffres, tandis que le ministre Laurent Fabius reste plus vague.

Dans les extraits suivants, le discours du ministre Jean-Marc Ayrault est construit à la voie active. Le sujet est essentiellement je, nous ou une entité humaine ou personnifiée. Par ailleurs, les verbes employés avec je sont en partie des verbes modaux ou semi-modaux. Dans le discours du secrétaire d'Etat, Harlem Désir, les phrases ont pour sujet le Brexit ou l'Europe et n'affichent pas de prises de positions claires.

Jean-Marc Ayrault, ministre des Affaires étrangères et du développement	*Harlem Désir, secrétaire d'Etat aux AE*
« **Je pense** à l'exemple de la COP21 qui a été enrichissant, sur le plan de la méthode. Ce succès, c'est la preuve que **nous sommes plus forts** à 28 que si la France avait négocié seule. […] » *Sénat, 28/06/2016.* « **Je ne sous-estime pas** l'impact du Brexit, qui pèse surtout sur le Royaume Uni. Mais **je profite de cette rencontre avec vous** pour vous faire une recommandation. » *New York, 22/09/16*	« Le départ du Royaume-Uni va en effet entraîner une série de renégociations au sein de l'Union européenne […] - il ne bénéficiera plus du fameux rabais mais l'UE ne bénéficiera plus non plus de sa contribution nette. » *Assemblée nationale 04/10/16* « L'Europe ne traverse pas aujourd'hui une crise de plus mais bien une crise du projet européen lui-même. » *Berlin, 21/10/16*

Cette opposition entre énonciation subjectivisée et énoncés objectifs affecte également les dénominations des personnalités et collaborateurs cités :

Laurent Fabius, ministre des AE et du développement international	*Annick Girardin, secrétaire d'Etat au Développement et à la Francophonie*
« C'est pourquoi je félicite et remercie particulièrement **mon collègue et ami Jean-Yves Le Drian** qui est à l'origine-accompagné par le Sénat- de cette rencontre. » *Paris, 14/10/15*	« […] la signature avec **le ministre de l'économie et des finances** du document cadre de partenariat entre la France et la Guinée » *Conakry, 13/09/14*
	Yamina Benguigui, ministre déléguée à la Francophonie
« **Mon ami, Michel Sapin**, a évoqué devant vous les travaux qui s'engagent au niveau international, et qui sont d'une très grande importance. » *Le Bourget, 04/12/15*	« La [DGLFLF], structure interministérielle rattachée **au ministre de la culture** et qui est à la disposition de **la ministre déléguée à la francophonie** […] » *Sénat, 30/01/14*

Ici, la secrétaire d'Etat A. Girardin, et la ministre déléguée Y. Benguigui nomment leurs collègues par leur fonction et leur statut. Le ministre L. Fabius, quant à lui, nomme ces mêmes collègues par leurs noms, et les qualifie par la relation qu'il entretient avec eux.
Ces différents exemples nous permettent de conclure que le discours des secrétaires d'Etat et des ministres délégués semblent donc plus objectifs, plus factuels et plus concrets, quand celui des ministres présentent plus de marques de subjectivité.

5. Conclusion

La construction d'un corpus tel que celui présenté ici permet d'établir des comparaisons entre différents paramètres : qui parle, quand, où, devant qui, à quelle occasion, etc. De ce fait, il nous est possible d'entreprendre des analyses concernant l'évolution du discours des représentants du MAE au cours du temps d'une part, mais aussi les variations dues aux conditions de production et d'énonciation de chaque prise de parole.
Dans cet article, nous avons pu observer que, lorsque l'on s'attache à la construction syntaxique du discours, les variations ne sont pas tant diachroniques que politiques, ou dues au statut du locuteur. Si un découpage du corpus par mandat montre que l'affiliation politique du chef du Gouvernement et/ou du représentant du MAE influe sur cette construction, une partition par locuteur nous permet d'affirmer que la position du locuteur influe également sur la syntaxe : les ministres présentent un discours plus verbal, vague et subjectif, tandis que les ministres délégués et secrétaires d'Etat se tournent vers des propositions plus nominales, précises et factuelles. Ces résultats pourront être approfondis par d'autres requêtes et d'autres comparaisons en croisant des paramètres tels que le genre ou le parcours antérieur de chaque locuteur, le lieu où le discours est initialement prononcé ou encore le public face auquel le locuteur s'exprime.

Bibliographie

Brunet É. (1981). *Le vocabulaire français de 1789 à nos jours : d'après les données du « Trésor de la Langue Française ».* Slatkine-Champion, 3 vol., p. 852, 518 et 454.
Charaudeau P. (2004) *Le discours politique. Les masques du pouvoir*, Paris, Vuibert, p. 255.
Duchastel J. (1993). « Discours et informatique : des objets sociologiques? », *Sociologie et sociétés*, 25 (2), p. 157-170.
Gobin C. et Deroubaix J.-C. (2010). « L'analyse du discours des organisations internationales. Un vaste champ encore peu exploré », *Mots. Les langages du politique* [En ligne], 94. URL : http://journals.openedition.org/mots/19872
Habert B. (2004). « Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs ». *Revue française de linguistique appliquée*, IX, 5-24. URL :

https://doi.org/10.3917/rfla.091.0005
Heiden S., Magué J.-P. et Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». *10th International Conference on the Statistical Analysis of Textual Data* - JADT 2010, Rome, Italie. p.1021-1032.
Mayaffre D. (2004). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République*. Paris : Champion, 292 p.
Oger C. et Ollivier-Yaniv C. (2006). « Conjurer le désordre discursif. Les procédés de « lissage » dans la fabrication du discours institutionnel », Mots. Les langages du politique, n° 81, p. 63-77

Classificazione manuale *vs* automatica dei resoconti delle chiamate ricevute da Telefono Amico Italia (2016 -2020)

Valentina Rizzoli[1], Andrea Azzoni[2], Giulia Rivellini[2], Arjuna Tuzzi[3]

[1]Sapienza Università di Roma– valentina.rizzoli@uniroma1.it
[2] Università Cattolica del Sacro Cuore di Milano –
andrea.azzoni02@icatt.it giulia.rivellini@unicatt.it
[3]Università di Padova– arjuna.tuzzi@unipd.it

Abstract
Telefono Amico Italia (TAI) provides a telephone help service for people in emotional distress. The operators summarize in short reports the contents of calls and a group of experts classifies them into categories. The present contribution moves from the analysis of the contents of the reports with a twofold aim of detailing the categories adopted by TAI and proposing an update. Two different automatic classification perspectives have been exploited: a top-down one which aims to assess the consistency of the contents of the phone calls with the categories adopted by TAI's experts, and a bottom-up one which seeks to obtain a classification of the texts by topic (topic extraction). The top-down approach, showed the goodness of the categories adopted by the experts, but also the need for a redefinition of some categories that showed large areas of semantic overlap. The bottom-up analysis, has both confirmed the presence of categories already adopted and allowed to recognize the existence of new emerging categories. The multivariate analysis of these new categories revealed a decisive role of the category *mental health* in monitoring suicide.

Keywords: helpline, content mapping, topic extraction, multivariate analysis

Riassunto
Telefono Amico Italia (TAI fornisce un servizio di aiuto telefonico per persone in stato di disagio emotivo. Gli operatori sintetizzano in schede i contenuti delle telefonate e un gruppo di esperti le classifica in categorie. Il presente lavoro muove dall'analisi dei contenuti delle schede con il duplice obiettivo di dettagliare le categorie adottate da TAI e proporne un aggiornamento. Due diverse prospettive di classificazione automatica sono state impiegate: una *top-down*, che ha come obiettivo valutare la coerenza dei contenuti con le categorie adottate dagli esperti di TAI, e una *bottom-up*, che cerca di ottenere una classificazione dei testi per tema (*topic extraction*). L'approccio *top-down* ha mostrato la bontà delle categorie adottate dagli esperti, ma anche la necessità di una ridefinizione di alcune categorie che hanno evidenziato ampie aree di sovrapposizione semantica.

L'analisi *bottom-up* ha sia confermato la presenza di categorie già adottate, sia permesso di riconoscere l'esistenza di nuove categorie emergenti. L'analisi multivariata di queste nuove categorie ha evidenziato un ruolo determinante della categoria *salute mentale* nel monitoraggio del suicidio.

Parole chiave: linea di assistenza, mappa dei contenuti, estrazione degli argomenti, analisi multivariata

1. Introduzione

Telefono Amico Italia (TAI) è un'organizzazione di volontariato che dal 1967 fornisce un servizio di aiuto telefonico a persone in stato di disagio emotivo. Nota e diffusa su tutto il territorio nazionale, TAI riceve annualmente circa 50.586 contatti telefonici cui rispondono operatori volontari, formati dall'associazione, con l'obiettivo di ascoltare con piena disponibilità e attenzione partecipata il racconto di chi chiama (l'appellante), cui viene garantito l'anonimato (*cfr.* Rivellini et al., 2015). Per ogni chiamata viene redatta una scheda, priva di dati sensibili, che raccoglie le caratteristiche dell'appellante e sintetizza il contenuto della telefonata. Le schede vengono poi analizzate e categorizzate da un gruppo di esperti che le classifica in categorie secondo una struttura che TAI utilizza da molti anni. Si tratta di 28 categorie (che rappresentano il "problema prevalente" della telefonata), raggruppate in 5 macro-categorie (*pratico*, *sessualità*, *emarginazione*, *relazioni*, *area del sé*), 6 se si considera la registrazione di un problema non emerso chiaramente (Figura 1).

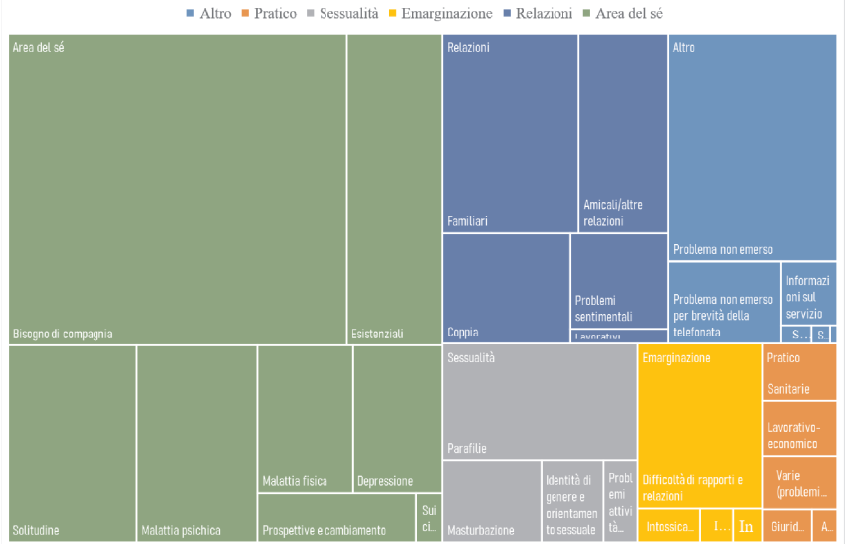


Figura 1. Macro-categorie assegnate da TAI e suddivisione nei rispettivi problemi prevalenti. Ogni area/colore è proporzionale alla presenza del problema prevalente/macro-categoria.

L'operazione di apporre l'etichetta del problema prevalente (scelta tra le 28 adottate da TAI) alla sintesi della telefonata è riconducibile, in senso lato, a un'analisi qualitativa e manuale del contenuto. Vale la pena sottolineare che classificare il testo e le parole che sono state utilizzate per rappresentare un problema della persona (obiettivo dell'operatore TAI) è diverso da individuare il disagio clinicamente (obiettivo di un terapeuta). Nel primo caso, infatti, viene classificato dall'operatore il problema "come descritto dalla persona" (ossia ne viene fornita una rappresentazione), senza effettuare una diagnosi (come nel secondo caso).
L'obiettivo del presente contributo è a) illustrare come individuare la rappresentazione del problema della persona a partire dai resoconti del turnista. A tal fine sono state adottate due diverse prospettive: 1) una *top-down*, che è stata svolta a partire dai testi suddivisi in base al problema prevalente e alla macro-categoria indicata nelle schede dagli operatori e che è principalmente finalizzata alla valutazione ed eventuale ridefinizione e rinnovo delle categorie preesistenti; 2) una *bottom-up*, che è stata svolta a partire dai testi suddivisi per anni e che è finalizzata a generarne eventualmente di nuove e osservare se cambiano o rimangono stabili negli anni. In più si intende: b) offrire una rappresentazione delle situazioni trattate da TAI a partire dall'analisi delle categorie in relazione alle caratteristiche degli appellanti.

2. Metodi

2.1. Costruzione del corpus

Il corpus messo a disposizione dal TAI è di grandi dimensioni (Tabella 1) ed è costituito da 203.316 schede contenti resoconti di telefonate relative a cinque anni (2016-2020). Ciascuna scheda contiene un resoconto lungo mediamente 117 parole e riporta alcune caratteristiche dell'appellante (evinte da chi riceve la chiamata): problema prevalente, genere, età, contesto relazionale, professione, provenienza, monitoraggio suicidio.

Tabella 1. Misure lessicometriche di base

Occorrenze – word token (N)	23.836.765
Forme grafiche – word type (V)	155.364
Rapporto tra forme grafiche e occorrenze (type/token ratio; V/N*100)	0,65
Percentuale di hapax (forme che compaiono una volta sola)	51,35

2.2. Analisi del testo delle schede

Al fine di osservare se i contenuti delle schede in termini di problema prevalente fossero coerenti con le categorie (procedura *top-down*), sono state individuate le parole che caratterizzano i problemi prevalenti secondo le categorie attualmente

adottate dagli operatori. È stata effettuata prima un'analisi delle corrispondenze lessicali sulla matrice parole × problema prevalente (Greenacre, 1984; Lebart et al., 1984) e poi è stato applicato il metodo Reinert (1986), implementato in Iramuteq (Ratinaud, 2009) ai 5 sub-corpora costituiti dalle schede della stessa macro-categoria di problema prevalente (*pratico*, *sessualità*, *emarginazione*, *relazioni*, *area del sé*), escludendo la categoria dei problemi non emersi. Il metodo Reinert è stato applicato anche ai sub-corpora costituiti dalle schede raggruppate per anno (2016, 2017, 2018, 2019, 2020) allo scopo di individuare gli argomenti prevalenti in modalità libera, cioè senza tenere conto delle categorie attribuite dagli operatori (procedura *bottom-up*). Questa analisi ha permesso di osservare in prospettiva diacronica l'emergere, lo scomparire o il modificarsi di alcune categorie di problema prevalente.

2.3. Analisi statistiche sugli appellanti

L'insieme dei dati è stato oggetto di analisi statistiche univariate, bivariate e multivariate, per una visione sintetica e facilmente interpretabile dei profili socio-demografici degli appellanti e dei loro problemi nel corso dei cinque anni considerati. L'analisi univariata è stata eseguita sulle tre differenti tipologie di categorie di problemi prevalenti: quella individuata come problema prevalente da TAI e le categorie ottenute tramite le analisi *top-down* e *bottom-up*. L'analisi bivariata è stata condotta condizionando la classificazione bottom-up alle caratteristiche socio-demografiche degli appellanti: genere, classe di età, provenienza, professione, occupazione e contesto relazionale, alle quali si aggiunge la variabile monitoraggio del suicidio, che registra la preoccupazione dell'operatore TAI riguardo la possibilità che l'appellante commetta suicidio. Attraverso un'analisi multivariata è stato costruito un modello predittivo per comprendere il monitoraggio del suicidio (Gareth *et al*., 2017). Sono state utilizzate come variabili indipendenti la classificazione *bottom-up* e le caratteristiche socio-demografiche degli appellanti. Come primo passo il dataset è stato suddiviso in training-set (per costruire il modello) e test-set (per la valutazione del modello). Dal momento che l'ampiezza delle classi del monitoraggio suicidio era pari a 98% per monitoraggio del suicidio = 0 (assente) e 2% per monitoraggio suicidio =1 (presente), è stato utilizzato il metodo di oversampling SMOOTE per bilanciare la classe del monitoraggio suicidio con la numerosità più bassa del training-set. Infine, sono stati implementati due modelli di classificazione: la Regressione Logistica (RL) e il Random Forest (RF). Il primo è diretto a stimare attraverso i coefficienti beta le variabili maggiormente influenzanti l'output. Il secondo si basa sulla costruzione di diversi *trees* i quali cercano di spiegare quali siano le variabili che impattano maggiormente il modello. Nell'output finale questo si coglie dal valore dell'importanza riconosciuta alle singole variabili.

3. Risultati

Dai risultati dell'analisi delle corrispondenze (*top-down*; Figura 2) è possibile osservare che i problemi prevalenti relativi alla macro-categoria *sessualità* occupano un quadrante a sé stante, a conferma che rappresenta tematiche poco sovrapponibili alle altre macro-categorie. Le macro-categorie *area del sé* e *pratico* risultano, viceversa, abbastanza sovrapponibili in termini di contenuti.

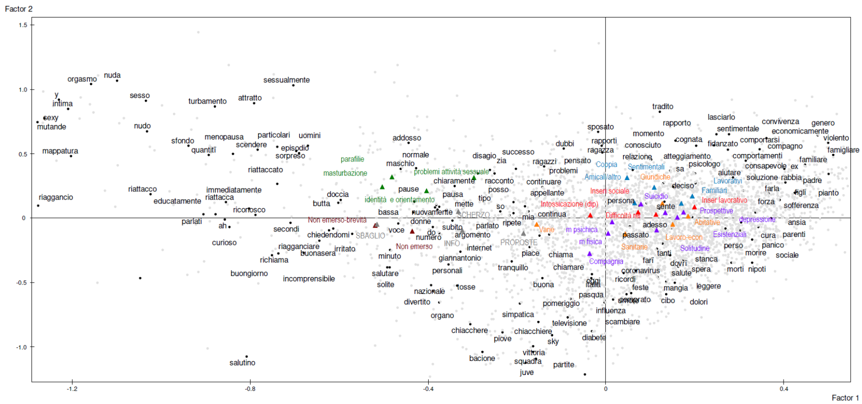


Figura 2. Primi due fattori dell'analisi delle corrispondenze parole × problema prevalente e macro-categorie (arancione-pratico, verde-sessualità, rosso-emarginazione, blu-relazioni, viola-area del sé).

Applicando il metodo Reinert al sub-corpus delle schede assegnate alla macro-categoria del problema *pratico* (*top-down*) si osserva che la classificazione automatica tende a raggruppare in un'unica classe chiamate relative ad appellanti specifici, ma, se confrontata con la classificazione manuale, i problemi individuati sono analoghi: di natura economica, salute fisica, ambito legale e giuridico. Sono inoltre visibili problemi legati a figure professionali specifiche. Rispetto alla macro-categoria *sessualità* la differenza più marcata è che l'analisi automatica individua due cluster differenti (se pur legati tra loro) relativi alle relazioni sessuali: uno più legato alle dinamiche di coppia (es. *rapporto sessuale, relazione*) e l'altro con maggiori riferimenti ai contesti di vita che intersecano le relazioni (es. *cena, suocero, scambista*). Il concetto di emarginazione (ulteriore macro-categoria) è legato a quello di relazione e per questa ragione alcuni contenuti dei cluster relativi alla *sessualità* sono presenti in altre macro-categorie (come evidenziato dall'analisi delle corrispondenze), soprattutto alla macro-categoria *relazioni*. La macro-categoria *area del sé* risulta la più vasta e varia. La classificazione automatica raggruppa le schede di alcuni appellanti specifici in cluster a sé stanti in quanto si riferiscono a temi ricorrenti o gli operatori riportano descrizioni molto simili per le telefonate ricorrenti degli stessi

appellanti. Emergono due cluster che rimandano a problemi prevalenti già presenti quali percorsi di vita (es. *vita, situazione, affrontare*) e salute mentale (es. *psichiatra, depressione*).
Dall'analisi *bottom-up* (metodo Reinert applicato ai sub-corpora relativi ai diversi anni) si osserva che la struttura delle classi individuate resta costante nei primi anni, mentre il numero aumenta nel 2020. I cluster sono stati confrontati e sulla base delle parole associate sono state generate nuove categorie relative al problema prevalente attribuibile alle schede (Tabella 2). Tra le nuove categorie emerse vi sono quelle relative alla *pandemia* Covid-19 (es. *coronavirus, uscire, contagio*) e situazioni particolari come *violenza in famiglia* (es. *violenza, famiglia, separare, madre, padre*), *lutto e dolore* (es. *dolore*, *disperato, sofferenza*). Nuovamente emergono i cluster (ricorrenti nel tempo) delle chiamate di alcuni appellanti specifici.

Tabella 2. Classi individuate bottom-up tramite il metodo Reinert e relativa presenza negli anni.

Categoria Bottom up	**2016**	**2017**	**2018**	**2019**	**2020**
Gestione politica della situazione pandemica					5,8
Questioni economico sociali	2,9	4,6	3,8	6,7	7,1
Lutto e dolore			8,4	6,8	4,7
Violenza in famiglia					3,9
Salute fisica			4,9		4,3
Covid - 19					6,4
Salute mentale	13,4	11,0	2,6	6,4	7,2
Rapporti-integrazione economico/lavorativo/relazionali	5,1	2,4	3,9	4,3	
Relazioni amicali/percorsi di vita	7,7	9,2	9,1	9,0	
Relazioni romantiche	7,0	9,8	11,4	6,1	4,7
Interruzione della telefonata				5,7	
Relazioni di coppia – sessuali				2,7	
Ricerca di dialogo					4,0
Sfogo su temi generici					6,2
Situazioni di vita specifiche	3,5	2,1	2,6	2,2	3,7
Solitudine legata alla mancanza di famiglia	11,4	9,8	3,4	3,6	
Bisogno di compagnia			10,1	9,7	3,9
Sessualità	8,9	6,6	11,6		3,4
Problematicità specifiche di un appellante	2,2	6,7			4,2
Ossessione urina					2,6
Identità di genere/orientamento sessuale	2,1				
Gestione della telefonata - linea/informazioni	5,5	6,3	10,6	2,6	
Gestione della telefonata - gestione rapporto appellante	14,0	7,9		9,9	6,8
Tentativi di approccio dell'appellante		4,0			2,3
Sport/noia - chiamate ricorrenti da appellanti specifici	10,3	6,8	8,0	7,4	5,4
Non classificati da software (pari a 23.555 schede)	6,1	12,8	9,7	17,1	13,2
Totale (%)	100,0	100,0	100,0	100,0	100,0

Dall'analisi univariata è emerso che le problematiche più frequenti riguardano l'*area del sé* e, in particolare, il *bisogno di compagnia* (Figura 1). Le donne sono costantemente più coinvolte in problemi legati alla solitudine, al bisogno di compagnia, alle relazioni romantiche e amicali e più in generale all'evolversi del percorso di vita. Tra gli uomini la sfera della sessualità è più ricorrente. La vita trascorsa con altri protegge dalla solitudine e dal bisogno di compagnia, ma non da una salute mentale compromessa. Tra gli anziani emergono i temi della solitudine, del lutto e del dolore e della pandemia Covid-19. Sia la RL che il RF hanno evidenziato come l'etichetta *salute mentale* sia una variabile di grande importanza per il monitoraggio del suicidio (nell' output del RFsi trova al secondo posto con valore della variabile importance pari a 0.045). Altre situazioni di criticità associate alla prevalenza di comportamenti che possono suggerire all'operatore la presenza di tendenze suicide riguardano le situazioni di *violenza domestica*, di *lutto e dolore*. L'accuratezza dei modelli previsivi è piuttosto elevata con un valore di *accuracy* pari all' 84% per la RL e del 98,5% del RF.

4. Discussione e conclusioni

La procedura *top-down* ha mostrato in linea di massima la bontà delle categorie utilizzate dagli operatori di TAI. Da un lato l'analisi delle corrispondenze ha permesso di osservare e valutare la coerenza tra etichetta e contenuto andando a individuare la relazione (più o meno sovrapponibili e quindi simili in termini di contenuti) tra le etichette. Dall'altro, il metodo Reinert ha permesso di individuare in modo automatico nuovi temi all'interno delle macro-categorie. Questa procedura ha permesso, inoltre, di fornire delle indicazioni utili nell'ottica di una eventuale ridefinizione di categorie preesistenti, per esempio aggiungendo dei problemi prevalenti all'area del sé, che risulta la più ampia e meno caratterizzata. La procedura *bottom-up* ha da un lato nuovamente confermato la bontà delle categorie; infatti, molte di quelle emerse tramite il metodo Reinert sono risultate analoghe ai problemi prevalenti già in uso dagli operatori. Dall'altro lato ha evidenziato l'emergere di nuove categorie che via via sono venute a delinearsi nel tempo, sia legate anche a nuove situazioni (pandemia Covid-19), che a problemi ricorrenti, come, per esempio, la gestione del lutto o la violenza. Grazie al contributo delle analisi multivariate è stato possibile capire, rispetto ad argomenti di grande interesse come il monitoraggio suicidio, quali fossero i problemi determinanti, evidenziando il ruolo preponderante, per esempio, della salute mentale. Trattandosi di un corpus composto da resoconti di telefonate (e non trascrizioni) non è stato possibile valutare la coerenza dell'etichetta apposta rispetto ai contenuti delle reali telefonate, ma solo rispetto ai resoconti. Il processo di fornire il resoconto può implicare una prima classificazione di contenuti. Questo lavoro contribuisce a una riflessione sul ruolo che svolge l'operatore chiamato a descrivere quanto esposto dall'appellante. I

risultati di queste analisi sono stati condivisi con TAI e, oltre a fungere da monitoraggio costante, sono stati presi in considerazione per la formazione degli operatori e per una revisione delle schede in essere. Future ulteriori collaborazioni con TAI potrebbero arricchire lo studio valutando anche le caratteristiche anagrafiche degli appellanti e i loro stati d'animo durante la telefonata.

Bibliografia

Gareth J., Witten D., Hastie T., Tibshirani R. (2017). *An Introduction to Statistical Learning with Applications in R.* Springer New York Heidelberg Dordrecht London.

Greenacre, M. J. (1984). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis; Correspondence Analysis and Related Techniques for Large Matrices*. New York: Wiley

Ratinaud P. (2009). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Computer software]. http://www.iramuteq.org

Reinert M. (1986). Un logiciel d'analyse lexicale: Alceste Les Cahiers d'Analyse des Donnees, 11(4), 471–481. http://www.numdam.org/item?id=CAD_1986__11_4_471_0

Rivellini G., Sironi E., Briccola D., and Vendemiati L. (2015). *L'ascolto: un approdo sicuro*. *Quando il telefono ci viene in aiuto*. Padova: Cleup.

How far is the risk? Detecting socio-psychological processes from consciousness to denial in social media exchanges

Valentina Rizzoli[1], Laura Soledad Norton[1],
Alessandro Meneghini[2], Mauro Sarrica[1]

[1]Sapienza Università di Roma – valentina.rizzoli@uniroma1.it
laura.norton@uniroma1.it mauro.sarrica@uniroma1.it
[2]Università di Padova – alessandro.meneghini.1@phd.unipd.it

Abstract
Within the framework of the computational social sciences, specifically in social psychology, one of the big issues is to automatically recognise socio-psychological processes in texts. There have been several attempts to identify such processes (e.g., the most well-known, emotions), although mostly with dictionary-based approaches and little still exists on the identification of psychological processes through machine learning, precisely because of the difficulty of the preliminary work required that is very specific to the application domain, namely the creation of a training set. The present contribution addresses the issue of operationalising socio-psychological processes as they are expressed in natural language to create categories for text classification through machine learning algorithms. In particular, within the framework of the online construction of risk (related to two global linked crises: climate change and the Covid-19 pandemic), it focuses on the implementation of a training set aimed at detecting processes which serve as barriers to implementing effective coping behaviours to face risks (i.e., moral justification, psychological distance, denial) in social media exchanges (i.e. Twitter). Problems and potentials of the adopted approach will be discussed, as well as the methodological and applicative implications of the study.

Keywords: linguistic features; machine learning; classification algorithms; socio-psychological processes; risk perception; global crises

1. Introduction

To automatically recognise socio-psychological processes in texts remains an open challenge in the field of computational social sciences and, in particular, social psychology. Such operation can be performed mainly through two different approaches: one dictionary-based (e.g., LIWC; Pennebaker et al., 2015) and one relying on machine learning. The former is the most often used one, but, although

it has proved its utility, it has limitations inherent in the approach itself, i.e. it does not allow to consider the context of the use of words, which is fundamental in identifying psychological processes. Even though the latter has been shown to be able to overcome these limits (cf. Hartman et al., 2019), it is still rarely used for the identification of socio-psychological processes, precisely because of the difficulty of the preliminary work required that is very specific to the application domain, namely the creation of a training set.
The present contribution proposes an attempt to operationalise some of the socio-psychological processes used to cope with crises. We examined how these strategies are expressed in natural language and created categories for text classification (i.e. the training set). In particular, we examined moral disengagement, distancing and denial in relation to environmental and health crises: climate change and the Covid-19 pandemic, which can be considered two among the various and intertwined typologies of risks we are facing as individuals and communities (Fuentes et al., 2020; Phillips et al., 2020).

1.1. Socio-psychological barriers to cope with risk-exposing events

Among the variety of socio-psychological barriers that prevent effective coping behaviours facing risk events, in this contribution we will focus on moral disengagement, distancing and denial processes since they are all processes that can build in communication and thus can find expression in natural language.
Bandura (2016), within the framework of the agentic perspective of social cognitive theory, defines *moral disengagement* as the mechanism through which people selectively disengage their moral self-sanctions from their harmful conduct. Several processes contribute to moral disengagement. *Moral, social, and economic justification* are used to attribute honourable purposes to the actors. The mechanism of *advantageous comparison* consists of framing an issue as the lesser of two evils, making it not only acceptable but even morally right. The *displacement of responsibility* is the obscuring or understating of the proper agentive role in causing harm. Finally, moral disengagement is reached by minimising or discrediting evidence of harm (*disregarding, minimizing, and disputing detrimental effects*). These processes have been described as problematic in relation to many issues, including climate change (Bandura, 2016).
The second barrier we consider in this paper is defined within the framework of the *Construal Level Theory* (CLT): the *psychological distance*. It is the perception of an object, event, or situation as not concretely and directly experienceable, thus distant from oneself (Trope & Liberman, 2010). An object can be perceived *geographically distant* when perceived as distant in space; *temporally distant,* when there is a perception of distance in time, either to the future or to the past; *socially distant*, when there is a belief that something may happen to others rather than to oneself; or *hypothetically distant*, when one doubts whether an event has

happened or may happen. Psychological distance has been shown to be a barrier to pro-environmental and health behaviours (Spence et al., 2012; Maiella et al., 2020; Blauza et al., 2020).
Third, we consider the process of *denial* (Cohen, 2001). It can be *literal*, that is the assumption that something did not happen or is not true; *interpretative*, when something is invested with a different meaning from what seems apparent to others; *implicatory*, when only psychological, political or moral implications are denied.
Drawing on this overarching framework, the present contribution aims to propose an automatic classification system of the processes above described. Specifically, the first objective of the work consists in their operationalisation, i.e. the identification of a univocal framework and classification system that includes all of them and serves as a guide for their identification in natural language. The second objective is to identify the typical linguistic characteristics of these processes (how they can be recognised in natural language). The third objective is to verify the effectiveness of classification for the detection of these processes.

2. Method

2.1. The corpora

Italian tweets related to the Covid-19 pandemic and climate change (selected on the basis of inherent hashtags, added over time based on those that were trending and related) have been collected through the rtweet R package (Kearney, 2019). The collection of the Tweets related to the Covid-19 Pandemic started at the end of January 2020, while as regards climate change, at the end of February 2022. They remain ongoing.

2.1.1. Operationalization of the socio-psychological processes

In order to generate useful categories for classification, we systematise the socio-psychological processes briefly presented (paragraph 1.1.), highlighting the overlapping and specificity of each of them (Table 1). Starting from the assumption that the downloaded texts concern the topic in question (i.e., Climate change or Covid-19 crisis) given the selection criterion, we individuated four categories positioned along a continuum from awareness to denial. At the first pole of the continuum, there is the *consciousness*, i.e. acknowledging (from a cognitive, emotional and value point of view) the presence of the risk in question. The *justification* category also implies risk awareness, but the no need of counteracting it is justified through different processes (i.e., moral, social, and economic justification, advantageous comparison, and implicatory denial and displacement of responsibility). In the *distance* category, the awareness remains, but if in the previous category the exposing-risk event concerned the person but was justified, now it is there but does not concern the person, it is considered far

away from oneself from various perspectives: temporal, geographical, and social. On the opposite side of the continuum with respect to consciousness, there is the *denial* that the object in question may exist or expose to risks. It can be literal (and coincides with the hypothetical distance) or interpretative.

Table 1. Summary of the categories for the classification and related socio-psychological processes

	SOCIO-PSYCHOLOGICAL PROCESSES		
	Psychological Distance (CLT)	***Denial***	***Moral Disengagement***
CATEGORY			
Consciousness			
Justification			Moral, Social, and Economic Justification
			Advantageous Comparison
		Implicatory Denial	Displacement of Responsibility
Distance	Temporal Distance		
	Geographical Distance		
	Social Distance		
Denial	Hypothetical Distance	Literal Denial	
		Interpretative Denial	Disregarding, Minimizing, and Disputing Detrimental Effects

2.2. Analyses

A selection of the collected tweets was analysed manually by three independent judges who debated the interpretation of the proposed model. Each tweet in which one of the processes was identified was labelled accordingly. Finally, 344 tweets for *Consciousness*, 117 tweets for *Justification*, 114 tweets for *Distance*, and 333 tweets for *Denial* were classified within each category, respectively, for a total of 908 tweets. The corpus constituting the training set was normalised by replacing uppercase with lowercase letters and lexicometric measures were calculated through TaLTaC2 software (Bolasco, 2010). With a total amount of word tokens of 24,833, a total amount of word types of 5,917, a types/tokens ratio of 23.83%, and an hapax percentage of 63.12%. Firstly, starting from manual analysis, recurring expressions in the natural language typical of each process related to the four categories were identified. This made it possible to

solidify and refine the conceptualisation of the classification model created. To test the effectiveness of the classification model, a machine learning algorithm for classification, which has already been used in similar applications (Fernandez et al., 2017), was applied to the training set. It consists of a corpus of 908 tweets divided into four categories (*Consciousness, Justification, Distance, Denial*). The decision to stop – initially - at the classification of the 4 categories and not to move on to the level of identifying the processes contained in them (which would be 9), although these have been identified by the manual analysis, lies in the dimensions that the training set has for now. Increasing the number of categories to be classified to 9 would have implied an insufficient number of samples for each category given the size of the texts in each of them. The training set underwent the AMNP procedure on R (Mikros & Perifanos, 2013). This procedure allows to sample the corpus in text chunks and assign them to their respective class based on the linguistic features. The whole corpus was split into text chunks of 200 words (i.e., the unit within which linguistic features are identified). As regards the linguistic features, the 100 most frequent words unigram, the 100 most frequent word bigrams, the 100 most frequent character bigrams and the 100 most frequent character trigrams were extracted from the text chunks and put in a matrix text chunks (117 rows) × linguistic features (401 columns), excluding values with variance close to 0 to reduce the noise within the classification process. Then Support-Vector Machines (SVM; Vapnik, 1995; Vapnik et al., 1997), which is a machine learning algorithm for classification were applied. SVM (discriminative classifiers) fit a margin-maximizing hyperplane between classes. It has been shown to be less prone to overfitting and generalising better (Hartman et al., 2019). The algorithms performed a 5 fold cross-validation, i.e., perform the training on 4/5 of the training set, using the remainder 1/5 as a validation set, and repeating the operation for a total of 5 times, each one with a different set of text chunks used as validation. Then, the final result of the cross-validation is the arithmetic mean between the results of the testing of each folder.

3. Results

3.1. Key features of the categories and socio-psychological processes

The manual analysis allowed us to individuate key distinctive features that characterise each category and the relative processes. The *consciousness* (C) category mainly contains descriptive tweets that state the situation, acknowledging the existence of the risk being discussed, for example, "*We are changing* the climate by deforesting the Amazon and polluting the sea irreversibly". In the *justification* (J) category we find tweets which express awareness of the risk, but are presented in a way where there's no need to counteract, with expressions like "rather than this one, focus on this other one"

are recurrent, as "these idiots then cry panic over the coronavirus. *It would be better* to control the deaths caused by pollution" (i.e., advantageous comparison). In the same category metaphors implying guilt and making "other" active agents are noted, as "the Chinese *plague spreader*" (i.e., displacement of responsibility). In the *distance* (Di) category, we find tweets where the risk is posed as far from the author, from different perspectives. From a spatial point of view, the distance is expressed with various geographical references that place the risk elsewhere, as "*The Philippines is particularly vulnerable* to climate change, or with temporal reference, that place the risk far in time because "there is no more/yet", as "But does the Covid emergency *still exist*?". Moreover, it is expressed mentioning other affected categories that make the event risky for someone else, as "Covid-19 *kills the elderly*". The *denial* (De) category contains expressions of literal denial or describing risk as something invented, as "the Climate Change *clowning*". Moreover, it contains expressions that re-interpret the situation "the truth is that..." "yes, but" or minimize or discredits through euphemisms, irony, etc. (i.e. interpretative denial). For example: "Some people *worry too much about normal flu*, […]. Evident coronavirus *paranoia*. In the USA everything is normal so far".

3.2. Classification through Machine Learning algorithm

To verify the effectiveness of the proposed classification we trained SVM algorithms. The confusion matrix obtained after the training showed that the accuracy in the classification task reaches the 76% (Table 2), with a precision of (C/De/Di/J) 0.82/0.91/0.13/0.059, recall of (C/De/Di/J) 0.87/0.78/0.091/0.33, and score of (C/De/Di/J) 0.85/0.84/0.11/0.10. As regards the categories of *Consciousness* and *Denial*, the majority of text chunks is correctly classified (Table 2). The *Justification* category is often misclassified as *Distance.* The algorithm fails to classify *Distance,* in fact, classifies it equally as *Distance* and *Consciousness* (Table 2).

Table 2. Support Vector Machine training

		Reference			
		C	**De**	**Di**	**J**
Prediction	**C**	7.4	0.6	1.0	0.4
	De	1.4	8.4	0.6	0.4
	Di	0.2	0.0	1.0	0.8
	J	0.0	0.2	0.0	1.0

4. Discussion and conclusions

This contribution constitutes the first attempt within an ongoing project to implement a training set to automatically classify socio-psychological processes related to risk perception. Although the corpus constituting the training set is too small, the training of the SVM algorithms showed quite satisfactory accuracy. However, the algorithm effectively predicts classification into two categories (*Consciousness* and *Denial*) out of four. The two categories are not only those that contain the most tweets but also those that are at the opposite poles of the model. Looking at the misclassifications, we see that many items classified as *distance* in the training set fall under the *justification* category during the training. Reasoning on the results of the key features identified with the manual analysis, we realised that some processes could be easily overlapped, based on such features. For example, the "Chinese plague spreader" (*displacement of responsibility – Justification* category) can also be interpreted as *geographical distance (distance* category*)*, since there is a reference to a Chinese virus/plague, leaving out the agentic term "*spreader*". This reasoning opens up a discussion of the limitations of classification algorithms concerning the need to classify this type of socio-psychological process: the fact that the categories must be mutually exclusive. On the other hand, the training set is not yet of satisfactory size, in terms of redundancy, and it is precisely the smaller categories that resulted to have less accuracy. The results, therefore, suggest a revision of the model and the definition of potentially ambiguous categories and emphasise the need for a larger corpus as a training set.

References

Bandura A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

Blauza S., Heuckmann B., Kremer K. and Büssing, A. G. (2021). Psychological distance towards COVID-19: Geographical and hypothetical distance predict attitudes and mediate knowledge. *Current Psychology*, 1-12.

Bolasco S. (2010). *Taltac 2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi.* Milano: LED.

Cohen S. (2001). *States of denial: Knowing about atrocities and suffering*. John Wiley & Sons.

Fuentes R., Galeotti M., Lanza A. and Manzano B. (2020). COVID-19 and climate change: a tale of two global problems. *Sustainability*, 12(20), 8560.

Hartmann J., Huppertz J., Schamp C. and Heitmann M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20-38.

Kearney M. W. (2019). rtweet: Collecting and analyzing Twitter data. Journal of Open Source Software, 4(42), 1829. doi: 10.21105/joss.01829, R package version 0.7.0.

Maiella R., La Malva P., Marchetti D., Pomarico E., Di Crosta A., Palumbo R. and Verrocchio M. C. (2020). The psychological distance and climate change: A systematic review on the mitigation and adaptation behaviors. *Frontiers in Psychology*, 11, 2459.

Mikros G. K. and Perifanos K. (2013). Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In Hovy, E., Markman, V., Martell, C. H. and Uthus, D. (eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext"*, 25-27 March 2013, Stanford, California. Palo Alto, California: AAAI Press, 17-23.

Norgaard K. M. (2011). *Living in denial: Climate change, emotions, and everyday life*. mit Press.

Pennebaker J.W., Booth R.J., Boyd R.L. and Francis M.E. (2015). *Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates* (www.LIWC.net).

Phillips C. A., Caldas A., Cleetus R., Dahl K. A., Declet-Barreto J. and Carlson C. J. (2020). Compound climate risks in the COVID-19 pandemic. *Nature Climate Change*, 10(7), 586-588.

Spence A., Poortinga W. and Pidgeon N. (2012). The psychological distance of climate change. *Risk Analysis: An International Journal*, 32(6), 957-972.

Trope Y., Liberman N. and Wakslak C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. *Journal of consumer psychology*, 17(2), 83-95.

Trope Y. and Liberman N. (2010). Construal-level theory of psychological distance. *Psychological review*, 117(2), 440.

Vapnik V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Vapnik V. N., Golowich S. and Smola A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In: Mozer, M., Jordan M. & Petsche T. (eds), *Advances in Neural Information Processing Systems* 9, pages 281–287, Cambridge, MA. MIT Press.

Life cycle of ideas in the *Journal of Personality and Social Psychology*: A history of US social psychology

Valentina Rizzoli[1], Matilde Trevisani[2], Arjuna Tuzzi[3]

[1]Sapienza University of Rome – valentina.rizzoli@uniroma1.it
[2]University of Trieste – matilde.trevisani@deams.units.it
[3]University of Padova – arjuna.tuzzi@unipd.it

Abstract
The present contribution offers a map of social psychology through the analysis of scientific publications in the field's pivotal North American journal, the *Journal of Personality and Social Psychology*. The titles of a comprehensive set of papers published in the journal (1965–2021) were collected, yielding a total of 10,222 items. The corpus thus constructed underwent several stages of preprocessing until the final conversion into a terms x time-points matrix, where terms are stemmed words and multi-words. After normalizing frequencies via a chi square-like transformation, clusters of words portraying similar temporal patterns were identified by functional (textual) data analysis and distance-based curve clustering. Among the best candidates in terms of the number of clusters, the solution with six clusters has been chosen and described, as it summarizes the evolution of the social psychology keywords (i.e., their life cycle) in an effective and interpretable manner. It reveals increasing, decreasing, and stable word trends, highlighting methods, theories, and topics that have become more popular in recent years, have lost popularity, or have remained in common use. Moreover, different trends with peaks in specific years highlight the most successful moments of theories and topics. The results point out the contribution that quantitative analysis of textual data can make to the study of a discipline's history.

Keywords: science mapping; history of social psychology; diachronic corpora; functional data analysis; curve clustering

1. Introduction

Each disciplinary history is constructed from a particular viewpoint and produced around a community (Danziger, 1995). A narrative of the history of ideas, that is, the evolution of a discipline in terms of theories, methods, and application domains over time, often becomes a history told in the light of what is known *ex post* and shared by mainstream schools and relevant scholars. This representation is only one of the possible ways of reconstructing and defining a discipline, which implies the risk of creating ceremonial or presentist histories (Hilgard et

al., 1991). Journal publications can serve as an alternative starting point for mapping the scientific production of a discipline since they reflect a history of the discipline built by the scientific community that gravitates around the journals themselves (Tuzzi, 2018). Moreover, since they represent a written and documented legacy, scientific journals offer a space for the exchange of ideas and dissemination of research results, as well as make available a record of scientific debate through time. The present contribution represents an attempt to reconstruct a disciplinary history from the history of unit-ideas, with all their mercurial and interdisciplinary character (Lovejoy, 1960). It offers a map of social psychology through the analysis of scientific publications in the field's pivotal North American journal, the *Journal of Personality and Social Psychology* (JPSP). Property of the American Psychological Association (which represents the largest psychology community in the United States and serves as a guiding reference for the discipline overseas as well), the JPSP represents a flagship journal for the discipline (Tesser, 1991; Cialdini, 2009).

2. Method

2.1. Corpus and preprocessing

All the titles (from the first publication in 1965 to the last in 2021) of the JPSP papers, that were entirely collected from Scopus database, comprise a total of 10,222 items. By means of TaLTaC2 software (Bolasco et al., 2000), the corpus thus collected has been tokenized and normalized, replacing uppercase with lowercase letters. Relevant multi-words with the number of occurrences ≥ 10 have been recognized and selected through an automatic information retrieval procedure that allows the identification of repeated informative sequences of words producing a compound (Pavone, 2010, 2018). Two encyclopedias of social psychology (Manstead et al., 1995; Baumeister and Vohs, 2007) and the index of keywords associated with each paper available in Scopus in the downloading process provided further multi-words (discipline-specific). After that, the list of corpus words and multi-words underwent a stemming procedure (Porter, 1980) and the stemmed words and multi-words thus produced have been considered together as final textual units. These two steps (multi-word identification and stemming) allowed optimizing the selection of meaningful keywords from the vocabulary.

2.2. Analysis

The analysis started from a term-document matrix that includes the most frequent (≥ 11) keywords (stemmed words and multi-words) by years (1,214 terms $\times$ 57 time-points). The procedure adopted for recognizing the dynamics of the keyword trajectories consists—from a statistical learning perspective—of three steps (cf. Trevisani and Tuzzi, 2015, 2018; Trevisani, 2018): (1) normalization of occurrences, (2) trajectory smoothing, and (3) grouping of words with similar trends

(curve clustering). Keyword frequencies have been normalized by a chi square-like transformation as this normalization previously proved effective in grouping words with similar timings of birth/death and presence/absence. Therefore, a functional data analysis approach was adopted. Each observation (i.e., the normalized frequency y_{ij} of a word i at a time-point j) was read as a realization of an underlying continuous function (i.e., a curve) that determines the word's development through time but is not directly observable. Steps 2 and 3 serve to extract from the set of curves groups of keywords that share a similar temporal evolution. First, B-spline smoothing is applied to reduce the irregular behavior of original trajectories. Then, the smoothed trajectories are grouped into clusters (curve clustering) by means of an algorithm (k-means) that uses the Euclidean distance between curves as a measure of similarity. The algorithm is repeated 20 times (varying the initialization) for each potential number of clusters (from 2 to 26), thus generating 20 possible partitions for each number. To determine the optimal number of clusters, about 50 different clustering quality criteria were made to compete against each other, and the best compromise among all the suggestions obtained was sought (Figure 1). Once the optimal number has been established, we identify among the 20 generated partitions those chosen by the quality criteria and, among the latter, the one that maximizes the degree of overlap with the others by computing an index (a generalized Rand index) representing cluster stability and consistency. Then, in order to choose the final clustering configuration, a qualitative reading is carried out by experts in social psychology to make sense of the clusters of words identified by the procedure and to start the interpretation of results.

3. Results

Among the best candidates in terms of the cluster number suggested by the basket of criteria (Figure 1), the solution with six clusters appears parsimonious and easy to read. Each cluster corresponds to a different trend that portrays the evolution of a set of keywords (Figure 2).

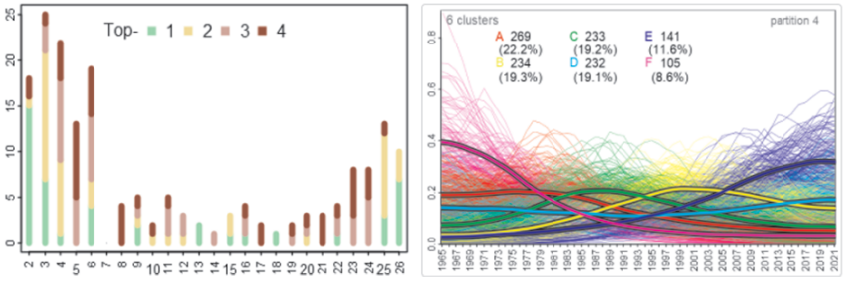


Figure 1. On the left, comparison between the choices of the optimal cluster number suggested by the basket of criteria as first (green), second (yellow), third (pink), or fourth (red) choice. On the right, solution with 6 clusters; the curve representing the average trend is highlighted and the number and percentage of words of each cluster are reported.

Cluster A (22.2% of keywords) includes keywords with a stable and slightly decreasing trend since the mid-1980s. The keywords it contains mainly refer to studies on attitude and behavior (*behavior, attitud*), aggression, and to the specific set-up of such studies, referring to the effects on participants subjected to stimuli (e.g., *exposur, rate, induc*) or assigned tasks to be completed. Other keywords concern socio-psychological processes and theories, such as social influence, causal attribution, and locus of control (*social influenc, causal attribt, locu of control*). Cluster B (19.3%) contains keywords with an increasing trend since the 2000s and stable afterwards. It includes references to the dominant perspective in social psychology, that is, social cognition (e.g., *social cognit, stereotyp, bia*), and some of its key themes, such as self and identity, emotions, intergroup relations, as well as theories (e.g., *self, ident, emot, social identit*) and fields of application (e.g., *health, work*). Cluster C (19.2%) contains keywords related to processes and methods particularly adopted in the 80s and 90s. Some examples are memory, decision-making, social judgment, social comparison (e.g., *memori, judgment, social judgment, social comparison*), but also the structure of personality and individual differences (e.g., *individu, individu differ, trait*). References to methods are mainly linked to self-report measures and measurement scales of constructs (e.g., *measur, scale, dimens, self report*). Cluster D (19.1%) includes keywords that have a nearly stable trend over time (i.e., evergreen topics). Some of these seem to be general words related to the description (e.g., *approach*) or conduct of studies (related methods), such as experiment (*experi*), control (*control*), prediction (*predict*), and correlation (*correl*). Other keywords refer to the groups (*group*, *group member*) and their dynamics, such as influence (*influenc*) and conflict (*conflict*). Cluster E (11.6%) contains keywords with an increasing trend: they become popular from the 2000s onward. These keywords recall theories such as the regulatory focus theory (*regulatori*), studies on personality (e.g., *big five, person trait, narciss*), culture (*cultur variat, cultur differ*), and intergroup processes (e.g., *outgroup*). Moreover, it contains references to recent popular themes, such as romantic relationships, well-being, life satisfaction (*romant relationship, romant partner, well be, life satisfact*), and some socially relevant issues, such as sexism and religion (e.g., *sexism, religi*). The cluster also includes method-related keywords, such as implicit measures, longitudinal studies, and meta-analyses (*implicit, longitudin, meta analysi*). Cluster F (8.6%) includes keywords that were prominent in the early years of the journal's life and then almost disappeared (decreasing trend). These mainly include words that refer to neo-behaviorism (e.g., *reinforc*), dissonance (*dissonanc*), group decision processes (e.g., *group discuss, conform, opinion chang*), game theories (e.g., *game, prison s dilemma*), as well as attitude and opinion change, conformism, and obedience to authority (e.g., *attitude chang, opinion chang, conform, complianc, authoritarian*). The word *negro* also

appears in this cluster, most likely because it has fallen into disuse in favor of politically correct language adopted afterward. The same happened with the word *subject,* which was dropped in favor of *participant.* Nevertheless, the sixth edition of the APA manual states that the word *subject* is again accepted in scientific language. With respect to the methods, this cluster contains keywords related to interviewing and field experiments (*intervie, field experi*).

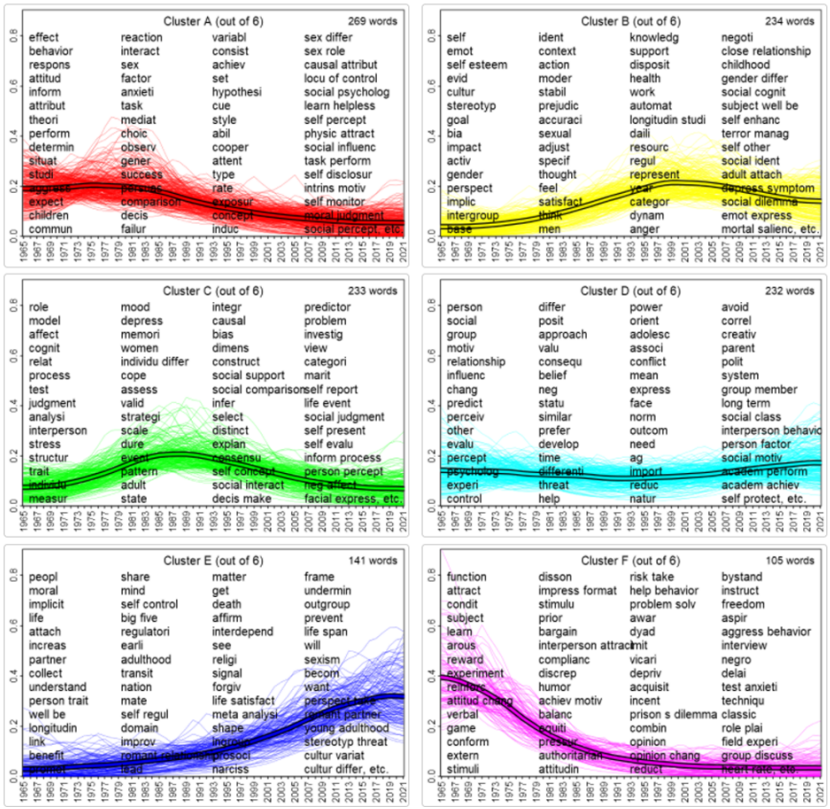


Figure 2. Representation of the trajectories of the words making the 6 clusters. The highlighted curve represents the average trend for the group

4. Discussion and conclusion

The present contribution aims to map the history of social psychology through the life cycle of keywords used in one of its major journals: the JPSP. Given the process of word selection that led to the identification of keywords and looking at those which fell into the same clusters, it was possible to trace the histories of theories, methods, and areas of application, on the one hand, but also the changes in language and therefore in the rules of the community within the discipline, on

the other. The analysis revealed increasing trends for keywords highlighting methods, theories, and topics that have become more popular in recent years (e.g., meta-analysis, regulatory focus theory, sexism, well-being); decreasing trends for words highlighting those that lost popularity (e.g., field experiment, neo-behaviorism, dissonance); stable trends for words highlighting evergreen methods and topics (e.g., correlation, conflict, group), and, finally, keywords that have remained in common use (e.g., social, change, predict). Moreover, different trends with peaks in specific years highlighted the most successful moments for theories and topics. For example, keywords related to the study of obedience to authority and conformism, as well as the legacy of the period of fascism/Nazism (Ross et al., 2010), emerged among those most in use in the early period of the considered time span. Again in the early years of the journal's life some processes, as attitude and opinion changes, exhibited peaks in popularity. This is the same period defined by McGuire (1986) as the era of attitude change, with the expected delay due to the timing of publication and the stabilization of concepts within the scientific production. Methods-related keywords revealed the dominant types of studies within the discipline in the mainstream tradition, namely, experimental and correlational studies. They also indicated those that have fallen into disuse (or, rather, have no place in mainstream publications), such as field experiments or interviews. Keywords with an increasing trend exhibited increased intertwining (confirming the choice of the journal's name) between social and personality psychology. Moreover, the constant presence of keywords related to group studies (even if with some related theories that have had different peaks of use) shows how the topic strongly characterizes social psychology.

This work has demonstrated the strengths of a knowledge-based system in detecting and grouping the life cycle of keywords and in drawing a historical map of a discipline. The goodness of result is also confirmed by comparison with further representations of the history traced by social psychology scholars and with further analyses for science mapping (e.g., topic modelling; cf. Rizzoli, 2018). Choosing a mainstream journal for the analysis allows tracing a mainstream narrative of the history of the discipline and individuating the areas with which it has been intertwined. However, this choice leaves out some currents and facets of the discipline as it mirrors only a particular viewpoint. Combining multiple journals might further enrich the discussion and discipline mapping (Rizzoli, 2018).

References

Baumeister R. F., and Vohs K. D. (2007). *Encyclopedia of social psychology*. Thousand Oaks, CA: Sage.

Bolasco S., Baiocchi F., and Morrone A. (2000). *TaLTaC2: Trattamento automatico*

Lessicale e Testuale per l'analisi del Contenuto di un Corpus [Computer software]. http://www.taltac.it/it/index.shtml
Cialdini R. B. (2009). We have to break up. *Perspectives on Psychological Science*, 4 (1), 5-6.
Danziger K. (1995). Neither science nor history? *Psychological Inquiry*, 6 (2), 115-117.
Hilgard E. R., Leary D. E. and McGuire G. R. (1991). The history of psychology: A survey and critical assessment. *Annual Review of Psychology*, 42(1), 79-107.
Manstead A. S., Hewstone M. E., Fiske S. T., Hogg M. A., Reis H. T. and Semin G. R. (1995). *The Blackwell Encyclopedia of Social Psychology*. Blackwell Reference/Blackwell Publishers.
McGuire W. J. (1986). The vicissitudes of attitudes and similar representations in twentieth century psychology. *European Journal of Social Psychology*, 16 (2), 89-130.
Pavone P. (2010). Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus. In S. Bolasco, I. Chiari, and L. Giuliano (Eds.) *Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, 9-11 June 2010, Sapienza University of Rome, pp. 131-140. LED.
Pavone P. (2018). Automatic Multiword Identification in a Specialist Corpus. In A. Tuzzi (Ed.), *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 151-166). Cham: Springer.
Porter M. F. (1980). An algorithm for suffix stripping. *Program*, 14 (3), 130-137.
Rizzoli V. (2018). Histories of Social Psychology in Europe and North America, as Seen from Research Topics in Two Key Journals. In A. Tuzzi (Ed.), *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 65-86). Cham: Springer.
Ross L., Lepper M. and Ward A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, and G. Lindzey (Eds.), *The Handbook of Social Psychology* (5th ed., Vol. 1, pp. 3-50). Hoboken, New Jersey: John Wiley and Sons.
Tesser A. (1991). Editorial. *Journal of Personality and Social Psychology*, 61 (3), 349-350.
Trevisani M. and Tuzzi A. (2015). A portrait of JASA: The History of Statistics through analysis of keyword counts in an early scientific journal. *Quality and Quantity*, 49 (3), 1287-1304.
Trevisani M. and Tuzzi A. (2018). Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based Systems*, 146, 129–141.
Trevisani M. (2018). Functional Data Analysis and Knowledge-Based Systems. In A. Tuzzi (Ed.), *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 167-187). Cham: Springer.
Tuzzi A. (Ed.). (2018). *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 167-187). Cham: Springer.

Analyse contrastive interlangue des marqueurs discursifs en Français Langue Étrangère et en Français Langue Maternelle

Minerva Rojas

Université Côte d'Azur – minerva.ROJAS@univ-cotedazur.fr

Abstract
The aim of this paper is to analyse the use of discourse markers (DMs) in oral production by French speakers and by learners of French as a foreign language in immersion. We created a monological and dialogical oral corpus (Corpus SCFLE) to calculate the proportion of DMs used individually and in both groups in relation to the number of words used. The results show that there is a remarkable variability across the two groups and that native speakers use more DMs than learners.

Keywords: discourse markers, corpus analysis, French as a Foreign Language, contrastive interlanguage analysis.

Résumé
Cette contribution a comme but d'analyser l'utilisation des marqueurs discursifs (MD) en production orale par des locuteurs francophones et par des apprenants de français langue étrangère en immersion. Nous avons créé un corpus oral monologique et dialogique (Corpus SCFLE) afin de calculer la proportion des MD utilisés individuellement et dans l'ensemble des deux groupes par rapport au nombre des mots utilisés. Les résultats montrent qu'il existe une variabilité remarquable dans l'ensemble des deux groupes et que les natifs utilisent davantage des MD par rapport aux apprenants.

Mots clés: marqueurs discursifs, étude de corpus, Français Langue Étrangère, analyse contrastive interlangue.

1. Introduction

Les marqueurs discursifs (MD) contribuent à la dynamique communicative ; ils servent à organiser les idées du locuteur et aident l'interlocuteur à interpréter les informations reçues (Martín Zorraquino & Portolés Lázaro, 1999). Pour les apprenants d'une langue étrangère l'utilisation des MD devient aussi essentielle car une absence ou une mauvaise utilisation rend difficile la communication dans la langue cible (Wierzbicka, 2003). Dans la recherche en acquisition des langues (RAL) l'analyse des MD est en relation avec l'étude des stratégies de communication (SC) et de la fluidité énonciative car ils contribuent à garder le flot

de la production orale sans interrompre la communication (Dörnyei & Scott, 1997 ; Dörnyei & Kormos, 1998) et leur absence ou leur excès relèvent des vulnérabilités de la fluidité énonciative (Segalowitz, 2010). Ainsi, les MD seraient un élément linguistique permettant de distinguer les apprenants des natifs ainsi que les apprenants les plus performants de ceux qui ne le sont pas (Hasselgren, 2002).

2. Cadre théorique

Notre intérêt par les MD s'ancre dans l'analyse des stratégies de communication (SC) en langue étrangère. Dans ce domaine, les MD, aussi nommés remplisseurs (*fillers*), serviraient à garder le canal de communication ouvert ; ils ne seraient pas reliés au sens du message mais à la gestion de la production pouvant aussi avoir des rôles discursifs (Dörnyei et Scott, 1997 : 193). La définition proposée dans ce cadre-là rejoint celles proposées par des spécialistes en MD. D'après Chanet (2004) les MD constituent des unités non référentielles et leur absence dans la production ne suppose aucune perte d'information essentielle à la compréhension du message. Leur utilisation a plutôt une fonction sémantico-pragmatique (Rodríguez Somolinos, 2011), et le sens exprimé diffère de celui associé aux catégories grammaticales classiques qu'ils recouvrent, comme les noms, les verbes ou les adjectifs (Dostie et Pusch, 2007).

Pour leur classification, nous avons combiné la proposition de Chanet (2004) et celle de Dörnyei et Scott (1997), en identifiant en tant que MD i) les connecteurs (*donc*, *alors*, *puis*, etc) ; ii) les particules recouvrant des catégories traditionnelles distinctes mais fonctionnant comme des MD à l'oral (*ben*, *voilà*, *du coup*, *genre*) ; iii) les mots phatiques (*tu vois*, *tu sais*, *hein*, *oui*, *ouais*), exclus par Chanet (2004) mais pris en compte par Dörnyei et Scott (1997). Enfin, il faut signaler que les pauses vocalisées *euh* ou *ah*, et les marqueurs régulateurs fournissant une rétro-alimentation communicative (*hm*, *d'accord*, *oui*) ont été exclus de cette classification.

Malgré l'intérêt croissant pour les MD en didactique des langues (Delahaie, 2011 ; Tran, Tutin et Cavalla, 2018), il y très peu d'études s'intéressant à l'utilisation des MD par des apprenants de français langue étrangère (FLE) (Deng, 2016 ; Sugiyama, 2017 ; Vladimirska, Gridina et Turlā-Pastare 2021) et à la comparaison entre apprenants et locuteurs natifs (Delahaie, 2009 ; Sugiyama, 2017). Par ailleurs nous n'avons pas trouvé de références concernant l'étude de leur emploi à long terme chez des apprenants de FLE. Pour toutes ces raisons, nous nous intéressons à l'analyse longitudinale de l'utilisation des MD en FLE et à l'analyse contrastive interlangue[1] (Granger, 2015) entre locuteurs francophones et apprenants de FLE. Le but est de savoir si leur utilisation diffère

[1] À la différence de l'analyse contrastive classique (analyse de deux langues différentes) l'analyse contrastive interlangue consiste à analyser la même langue, dans notre cas le français, en différents moments de son apprentissage (Granger, 2015).

quantitative et qualitativement entre le français langue maternelle et le FLE et comment leur utilisation évolue quantitativement chez les apprenants.

3. Méthodologie

Pour l'analyse des MD nous avons exploré le corpus oral annoté de français Corpus SCFLE composé de 47.025 mots. Ce corpus a été constitué dans le cadre d'une recherche doctorale entre les années 2015-2018 à l'Université Savoie Mont Blanc (Rojas Madrazo, 2020).

3.1. Participants et tâches

Les participants (N=26) sont divisés en deux groupes. L'échantillon d'apprenants (N=12) est constitué par tous les étudiants de L.E.A. de la mention anglais-FLE[2] promotion 2015-2018 de l'Université Savoie Mont Blanc. Le groupe des locuteurs francophones (N=14) est composé par des étudiants de la même licence, L.E.A., mention espagnol-anglais inscrits dans la même université ; l'échantillonnage des participants a été non-aléatoire basé sur le volontariat.

Tous les participants ont effectué les mêmes tâches, dont quatre en monologue et quatre en interaction. Les tâches en monologue consistent à regarder une courte séquence vidéo et à raconter, sans temps de préparation, la scène présentée dans la vidéo ; ce type de tâches (*story retell*) est souvent utilisé dans la RAL. Quant aux tâches en interaction, elles se déroulent en binômes et consistent à parler des thèmes de communication reliés à ceux proposés par le CECRL (2001) tels que les applications mobiles, la cuisine, la télévision et les moyens de transport, déclenchées par un texte introductoire et des questions sur le thème concerné.

Le recueil des productions des apprenants a été organisé en quatre collectes de données étalées entre la première année et la troisième année de licence : novembre 2015 (t1), mars 2016 (t2), novembre 2016 (t3) et novembre 2017 (t4). Dans chacune des collectes de données, les apprenants ont effectué une tâche en monologue et une tâche en interaction. En ce qui concerne le groupe de natifs, les tâches ont été effectuées entre novembre et décembre 2015 en plusieurs séances selon la disponibilité des étudiants.

3.2. Transcription et annotation du corpus

Pour la création du corpus nous avons utilisé la suite *EXMARaLDA* (Schmidt et Wörner, 2009) qui compte sur trois logiciels indépendants : *Partitur Editor*, *Coma* et *Exakt*. Dans le *Partitur Editor*, utilisé pour transcrire les productions orales enregistrées, trois couches ou pistes de transcription ont été créées dont la principale consacrée à la transcription orthographique des productions orales suivant les normes de transcription CHAT (MacWhinney, 2000). Ainsi, les MD

[2] Les étudiants apprenant le FLE sont des étudiants étrangers installés en France pour effectuer leurs études supérieures en Langues Étrangères Appliquées (L.E.A.)

sont signalés dans la piste principale précédés par deux virgules (*,,ben*).

3.3. Analyses

À partir de la collecte des données brutes (nombre de MD utilisés par chacun des locuteurs et nombre des mots de chaque production), nous avons calculé un taux individuel d'utilisation de MD. Le taux est calculé en divisant le total des MD utilisés par le nombre des mots de la production observée, multiplié par 100. Ensuite, les résultats ont été soumis à des statistiques descriptives pour mesurer les tendances centrales et la variabilité de chaque groupe (francophones et apprenants). Le test U de Mann-Whitney pour des échantillons non appariés a été lancé afin de savoir si les éventuelles différences entre les groupes sont significatives statistiquement. En outre, les dix MD les plus utilisés ont été extraits afin d'analyser leur utilisation en calculant leur taux par rapport au nombre total des mots des corpus[3] et comparer les choix des deux groupes dans chacune des tâches.

Concernant l'évolution du taux des MD chez les apprenants, nous avons lancé deux tests : le test Friedman pour 4 échantillons appariés pour analyser les différences entre les quatre collectes de données ; et le test Wilcoxon pour 2 échantillons appariés pour analyser les différences entre la première et la quatrième collecte de données. Enfin, les taux individuels des MD ont été soumis à des calculs de corrélation de Spearman (r_S) pour examiner si la relation entre le taux des MD et le temps est directe ou inversement proportionnelle et si elle est significative statistiquement (par des raisons d'espace nous ne rapporterons pas ces résultats).

4. Résultats

4.1. Taux de MD pour 100 mots. Différences entre les deux groupes

Tableau 1. Taux de MD dans les deux groupes. Résultats test U de Mann-Whitney.

	apprenants n=12					natifs n=14					U-MW	W	Z	Sig. Asint. (2-q)
	M	*Mdn*	*SD*	*Min*	*Max*	*M*	*Mdn*	*SD*	*Min*	*Max*				
Taux MD t1 mon.	4.29	3.320	2.86	1.43	11.69	4.90	4.414	2.75	1.53	10.00	67.00	145.00	-.87	.382
Taux MD t2 mon.	3.31	1.894	3.15	1.25	10.48	5.60	5.175	2.67	1.19	10.83	42.00	120.00	-2.16	.031*
Taux MD t3 mon.	3.96	3.267	2.28	1.58	10.23	5.64	5.861	2.11	2.08	8.70	43.00	121.00	-2.11	.035*
Taux MD t4 mon.	3.38	3.465	1.63	.91	7.41	4.59	3.986	2.09	2.21	9.21	47.50	125.50	-1.88	.060
Taux MD t1 int.	3.86	3.715	1.76	1.14	7.33	5.28	4.905	3.23	.89	11.25	64.00	142.00	-1.03	.304
Taux MD t2 int.	4.14	3.537	1.66	1.71	6.88	5.14	4.508	2.25	2.10	10.78	58.00	136.00	-1.34	.181
Taux MD t3 int.	3.33	2.750	1.87	.92	6.74	6.22	4.135	4.46	2.83	17.65	45.00	123.00	-2.01	.045*
Taux MD t4 int.	4.34	3.985	2.09	1.18	7.29	5.10	4.898	2.24	1.05	10.60	70.00	148.00	-.72	.471

[2] Le nombre de mots du corpus SCFLE est distribué ainsi : le corpus des apprenants contient 22.397 mots (monologue = 6.785 ; interaction = 15.612) ; le corpus des natifs comprend 24.628 (monologue = 8.767 ; interaction = 15.861). Le temps total transcrit et annoté du corpus SCFLE est de 284 minutes dont 97,096 minutes de monologue (53,831 minutes pour les apprenants ; 43,264 minutes pour les natifs) et 186,775 minutes d'interaction (110,425 minutes pour les apprenants ; 76,350 pour les natifs).

D'après les résultats descriptifs (tableau 1), tant dans le groupe de natifs que dans le groupe d'apprenants, il existe une variabilité nette entre les participants, constatée par les valeurs de l'écart-type (*SD*) et les valeurs minimales (*Min*) et maximales (*Max*). Malgré la variabilité des résultats, les locuteurs francophones utilisent davantage les MD ; les valeurs de la moyenne (*M*) et de la médiane (*Mdn*) obtenues par les locuteurs francophones restent au-dessus des résultats du groupe d'apprenants tout au long de l'étude dans les deux types de tâche (monologue et interaction). Les résultats du test U de Mann-Whitney montrent que l'utilisation des MD en monologue est significativement différente entre les deux groupes dans les collectes de données intermédiaires ((t1) $z = -.87$, $p > 0.05$; (t2) $z = -2.16$, $p < 0.05$; (t3) $z = -2.11$, $p < 0.05$; (t4) $z = -1.88$, $p > 0.05$). Dans la tâche en interaction, il n'existe pas de différences significatives sauf dans la deuxième collecte de données ((t1) $z = -1.03$, $p > 0.05$; (t2) $z = -1.34$, $p > 0.05$; (t3) $z = -2.01$, $p < 0.05$; (t4) $z = -.72$, $p > 0.05$).

En ce qui concerne les taux des MD extraits, les différences constatées portent sur certains choix (figure 1) ; les natifs utilisent davantage *du coup* et *enfin*

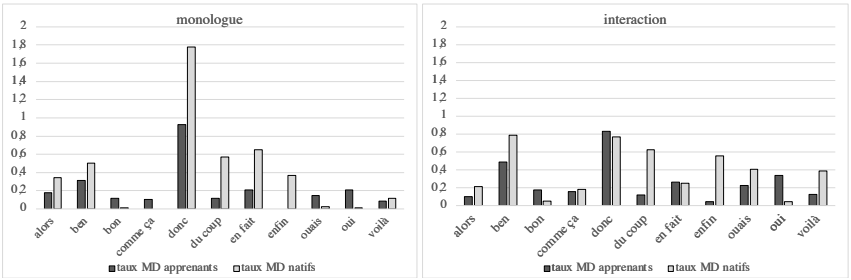


Figure 1. Différences apprenants - francophones. Taux des MD extraits pour 100 mots.

(monologue et interaction), *en fait* (monologue), *genre* et *voilà* (interaction). En revanche, les apprenants préfèrent *ouais*, *oui*, *bon*, *comme ça* (monologue) et *oui* (interaction). Des différences plus équilibrées surviennent dans les cas du MD *alors* (interaction et monologue), *en fait*, *comme ça* et *donc* (interaction). Il faut enfin signaler que *donc* est le MD privilégié par les deux groupes, mais il est plus utilisé par les locuteurs francophones en production monologique.

4.2. Taux de MD. Évolution du groupe d'apprenants

Les résultats descriptifs du taux des MD (tableau 1) montrent une diminution des valeurs de la moyenne d'un point entre le début et la fin de l'étude dans la tâche monologique. Cette diminution est constatée également dans les valeurs de l'écart-type, ce qui indiquerait que les différences entre les apprenants se resserrent à la fin de l'étude. D'ailleurs, l'étendu entre les valeurs minimales et les valeurs maximales diminue du début à la fin de l'étude à cause de la

diminution des valeurs maximales lors de la dernière collecte de données (t4). Dans la tâche en interaction, les résultats restent proches des valeurs de la tâche en monologue, mais les valeurs de la moyenne et de l'écart type augmentent légèrement du début à la fin. La variabilité entre les apprenants est moins marquée en interaction mais elle est aussi notable. D'après les résultats du test Friedman, il n'existe pas de différence significative entre le taux de MD des 4 collectes de données ni dans la tâche en monologue ($Q = 3.60$, $p > 0.05$) ni dans la tâche en interaction ($Q = 3.10$, $p > 0.05$) ; ce qui est corroboré par les résultats du test Wilcoxon tant dans la tâche en monologue ($Z = -.71$, $p > 0.05$) que dans la tâche en interaction ($z = -.78$, $p > 0.05$).

Tableau 2. Résultats test Friedman ; résultats test Wilcoxon.

Test Friedman Taux MD (t1 - t2 - t3 - t4)	N	Q	Sig. Asint. (2-queues)
Taux MD monologue	12	3.60	.308
Taux MD interaction	12	3.10	.376
Test Wilcoxon Taux MD (t1-t4)	**N**	**Z**	**Sig. Asint. (2-queues)**
Taux MD monologue	12	-.71	.480
Taux MD interaction	12	-.78	.433

5. Conclusion

Dans cette étude nous avons constaté que les natifs utilisent une proportion plus élevée de MD en monologue et en interaction. Pourtant les différences ne sont significatives que dans les collectes des données intermédiaires ce qui nous fait suggérer qu'elles ne sont pas en relation avec le temps d'immersion des apprenants. Nous avons également remarqué que les différences entre les apprenants et les natifs relèvent plutôt du choix des MD utilisés ; chaque groupe a des MD privilégiés et les natifs en utilisent un répertoire plus large. Quant à l'évolution du groupe d'apprenants, les résultats montrent que l'utilisation de MD varie très peu et les changements survenus ne sont pas significatifs statistiquement. Pourtant il existe une grande variabilité entre les apprenants dans la proportion de MD utilisés du début à la fin de l'étude. Il serait donc pertinent d'analyser les trajectoires individuelles et les choix des apprenants pour comprendre au mieux l'utilisation de MD à long terme et les différences entre les sujets.

Bibliographie

Casteele A.V. (2011) El uso de los corpora en el análisis de los marcadores discursivos por aprendientes de ELE. In C. Hernández González, A. Carrasco Santana, E. Alvarez Ramos (dir.) *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como*

lengua extranjera, pp. 527-534. ASELE & Universidad de Valladolid.
Chanet C. (2004). Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. *Recherches sur le français parlé,* 18: 83-107.
Delahaie J. (2009). Oui, voilà ou d'accord ? Enseigner les marqueurs d'accord en classe de FLE. *Synergies pays scandinaves*, 4: 17-34.
Delahaie J. (2011). Les marqueurs discursifs, un objet d'enseignement pertinent pour les étudiants Erasmus ?. *Éla. Études de linguistique appliquée*, 162: 153-163.
Deng D. (2016). Oui, voilà : analyse des deux marqueurs discursifs utilisés par les locuteurs du français d'origine chinoise en France. *Cahiers AFLS* 20(1): 43-69.
Dörnyei Z. and Scott M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language learning*, *47*(1): 173-210.
Dörnyei Z. and Kormos J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in second language acquisition*, 20(3): 349-385.
Dostie G. and Pusch C. D. (2007). Présentation. Les marqueurs discursifs. Sens et variation. *Langue française*, 2: 3-12.
Granger S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1): 7-24.
Hasselgren A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. Dans S. Granger, J. Hung & S. Petch-Tyson. *Computer learner corpora, second language acquisition and foreign language teaching*, pp. 143-174. John Benjamins.
Martín Zorraquino M.A. and Portolés Lázaro J. (1999). Los marcadores del discurso. *Gramática Descriptiva de la Lengua Española. Tercera parte. Entre la oración y el discurso.* Espasa Calpe, pp. 4051-4213.
Rojas Madrazo M. (2020) *Stratégies de communication, fluidité et lexique en production orale. Étude longitudinale d'apprenants en immersion*. Thèse de doctorat. Université Savoie Mont Blanc.
Segalowitz N. (2010). *Cognitive bases of second language fluency*. Routledge.
Schmidt T. and Wörner K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, *19*(4): 565-582.
Sugiyama K. (2017). Différences dans l'utilisation des marqueurs discursifs : analyse comparative entre apprenants et Français natifs. In H. Tyne, *Le Français en contextes*. [En ligne]. Presses universitaires de Perpignan. doi :10.4000/books.pupvd.2880
Tran T.T.H., Tutin A. and Cavalla C. (2016). Pour un enseignement systématique des marqueurs discursifs à l'aide de corpus en classe de FLE: l'exemple des marqueurs de reformulation. *Linguistik Online*, 78(4): 113-128.
Vladimirska E., Gridina J. and Turlā-Pastare D. (2021). Les marqueurs discursifs du français : un regard croisé sur une catégorie controversée. *Kalbotyra,* 74: 268–285.
Wierzbicka A. (2003) *Cross-cultural pragmatics: the semantics of human interaction,* Mouton de Gruyter.

Potenzialità di TaLTaC nella anonimizzazione di sentenze della magistratura

Maria Francesca Romano[1], Giovanni De Gasperis[2], Pasquale Pavone[1], Sergio Bolasco[3]

[1] Institute of Economics &EMbeDS-Scuola Superiore Sant'Anna – m.romano@santannapisa.it p.pavone@santannapisa.it
[2]Università degli Studi dell'Aquila & EMbeDS – giovanni.degasperis@univaq.it
[3]Università Roma La Sapienza & EMbeDS – sergio.bolasco@uniroma1.it

Abstract
Documents originating from the Public Administration are an important source of data for the analysis of social and economic phenomena, but the presence in the text of references to individuals make it necessary to have anonymization procedures in order to comply with privacy regulations.
References to individuals are of a different nature, depending also on the type of documents available. In the sentences of the criminal judiciary, for example, the individuals involved have very different roles (defendant, offended party, civil party, witness, etc.) and are identified by their name, place and date of birth but also with the tax code for the sentences issued in more recent times. In the case of civil judgments of divorce, moreover, among the physical persons mentioned there may also be children.
A further guarantee of *privacy* would impose to make anonymous the quotations of professional figures (lawyers, experts, and the same judges) included in any judgment both civil and criminal.

Keywords: TaLTaC, anonymization, legal texts, privacy, text mining

Riassunto
I documenti provenienti dalla Pubblica Amministrazione sono una importante fonte di dati per l'analisi di fenomeni sociali ed economici, ma la presenza nel testo di riferimenti a persone fisiche rendono necessario disporre di procedure di anonimizzazione per rispettare le norme relative alla *privacy*.
I riferimenti alle persone fisiche sono di diversa natura, dipendendo anche dal tipo di documenti disponibili. Nelle sentenze della magistratura penale, per esempio, le persone fisiche coinvolte hanno ruoli anche molto diversi (imputato, parte offesa, parte civile, testimone, etc) e sono identificate da nome e cognome, luogo e data di nascita ma anche

con il codice fiscale per le sentenze emesse in tempi più recenti. Nel caso di sentenze civili di divorzio, inoltre, tra le persone fisiche citate possono trovarsi anche eventuali figli, anche di età minore.
Un'ulteriore garanzia di *privacy* imporrebbe di rendere anonime anche le citazioni delle figure professionali (avvocati, esperti, e gli stessi giudici) inserite in qualsiasi sentenza sia civile che penale.

Parole chiave: TaLTaC; anonimizzazione; testi giuridici; *privacy*; *text mining*

1. Introduzione

I documenti provenienti dalla Pubblica Amministrazione sono una importante fonte di dati per l'analisi di fenomeni sociali ed economici, ma si scontrano con la presenza nel testo di riferimenti a persone fisiche che rendono necessario disporre di procedure di anonimizzazione per rispettare le norme relative alla privacy (Ohm, 2009).
L'anonimizzazione è di particolare rilevanza quando i documenti sono sentenze della magistratura (Csányi *et al*, 2021; Glaser *et al*, 2021; Povlsen *et al*, 2016). Nel caso di sentenze penali, le persone fisiche citate hanno un numero di informazioni differenti a seconda del loro ruolo: gli imputati sono identificati da nome e cognome, luogo e data di nascita, mentre minori informazioni sono presenti per le parti civili, testimoni o figure professionali come gli avvocati, i periti e gli stessi magistrati. Va considerato inoltre che nelle sentenze civili si possono trovare informazioni anche su minori, come nel caso di sentenze di divorzio.
Le sentenze, pur rispettando delle regole di scrittura generali, sono soggette ad una redazione libera. Nel lavoro proponiamo una procedura di *text mining* finalizzata all'individuazione di tutte le informazioni che possono violare le norme della *privacy*, permettendo di individuare anche quelle scritte in forme non corrette. La finalità è quella, inoltre, di restituire i testi in forma anonimizzata, garantendone l'analisi ai ricercatori interessati.

2. Elementi da oscurare

Nelle sentenze emesse dalla magistratura sono presenti numerosi elementi che necessitano di essere oscurati. Nella Tabella 1 riassumiamo brevemente tali elementi, suddividendo tra i tipi di oscuramento indispensabili e quelli accessori:

Tabella 1. Elementi presenti nelle sentenze da oscurare

Elemento	Ruolo nella sentenza	Tipo
Codice fiscale	Imputati, ricorrenti, testimoni, periti, minori, parti civili	indispensabile
	Avvocati, giudici, presidenti della corte	accessorio
Nome e cognome	Imputati, ricorrenti, testimoni, periti, parti civili, minori	indispensabile
	Avvocati, giudici, presidenti della corte	accessorio
Indirizzi residenza o domicilio	Imputati, ricorrenti, testimoni, periti	indispensabile
	Avvocati	accessorio
Data di nascita	Imputati, ricorrenti, minori	indispensabile
Luogo di nascita	Imputati, ricorrenti, minori	consigliabile
Ente giuridico	Diretto (imputato, parte civile)	consigliabile
	Indiretto (datori di lavoro)	accessorio
Partita IVA / codice fiscale	Enti giuridici privati / pubblici	indispensabile

Un caso particolare è quello della presenza di enti giuridici, sia pubblici che privati: questi possono avere ruolo diretto (come imputati, parti civili, o periti) oppure indiretto (in quanto datori di lavoro di persone fisiche con ruolo diretto). In entrambi i casi è comunque necessario identificare gli enti giuridici citati nelle sentenze e l'eventuale Partita IVA o altro codice che renda possibile un loro riconoscimento.

3. Metodi utilizzati

Il lavoro che presentiamo illustra una macro-procedura realizzata con il software TaLTaC (Bolasco, 2010; Bolasco e De Gasperis, 2017) con l'obiettivo di ricostruire un Corpus anonimizzato.

Le risorse linguistiche già disponibili nel software permettono il riconoscimento semantico di nomi propri e luoghi geografici, nonché la categorizzazione di figure parentali (Bolasco *et al*, 2007); le procedure utilizzate e sperimentate in altri lavori scientifici (Bolasco, 2005; Romano *et al*, 2020) possono poi permettere il riconoscimento di enti giuridici, sia pubblici che privati.

Ulteriori procedure sono state definite per identificare la successione di forme grafiche, secondo differenti sequenze di scrittura del codice fiscale.

Dopo la prima lettura del Corpus con TaLTaC è stata eseguita la Normalizzazione, salvaguardando le forme grafiche in maiuscolo (attribuendo categoria grammaticale MAJ anche alle forme miste).

Nella Tabella 2 vengono sinteticamente descritte le azioni effettuate per l'attribuzione della categoria semantica (CATSEM).

Tabella 2. Elementi presenti nelle sentenze da oscurare

Codice fiscale	codici fiscali scritti correttamente:	FG di lunghezza 16 e MAJ	Attribuita CATSEM CodFisc
	codici fiscali non scritti correttamente	Max due caratteri mancante o ripetuto	Attribuita CATSEM CFerr
	codice fiscale scritto in parti separate ma contigue	Lessicalizzate le MW identificate	Attribuita CATSEM CFerr
	codici fiscali degli avvocati	CATSEM(Titoli) LAG6 (CATSEM(CodFisc) OR CATSEM(CFerr))	Attribuita CATSEM AvvCF
Indirizzo di domicilio / residenza	attribuita CATSEM=preindirizzo alle FG quali (via, viale, corso, piazza, piazzale, strada, vicolo, salita, passaggio, etc)	CATSEM(preindirizzo) LAG6 CATGR(NUM)"	Attribuita CATSEM(INDIRIZZO)
Data di nascita e luogo di nascita	Identificate le sequenze possibili del tipo: "n(at?) (*) LUOGO (*) DATA" "n(at?) (*) DATA (*) LUOGO"		Attribuita CATSEM(ANAG)
Nomi	I nomi propri sono riconosciuti da TaLTaC con CAT(NM).		Attribuita CATSEM(NOME)
Cognomi	Individuate le sequenze di FG con CAT=NM e/o CATGR=MAJ Il risultato della query è stato importato in un nuovo corpus. attribuendo CATSEM(COGNOME) a tutte le forme grafiche non riconosciute come NM. La lista delle FG con CATSEM=COGNOME diventa poi l'input per l'attribuzione della CATSEM(COGNOME) nel Corpus originario.		Attribuita CATSEM(COGNOME)
Ente giuridico	Individuate la sequenza di FG con CATGR(MAJ) che precedono o seguono le FG già individuate come forma societaria (spa, srl, cooperativa, etc).		Attribuita CATSEM (ENTEG)

Riesportazione del Corpus anonimizzato
Viene adoperata la procedura prevista in TaLTaC per sostituire/correggere errori ortografici. In questa sede si sono oscurati tutti gli elementi qui considerati nel modo seguente ottenendo un Corpus anonimizzato. Nel nuovo Corpus anonimizzato gli elementi da oscurare compaiono quindi con l'indicazione della categoria semantica attribuita.
Controllo della correttezza delle procedure
Il controllo è stato effettuato in due step:
1) conteggio degli elementi oscurati per sentenza;
2) estrazione casuale di un numero limitato di sentenze per una lettura diretta, per calcolare il numero di elementi correttamente oscurati, i falsi negativi (elementi non anonimizzati) ed i falsi positivi (elementi che non dovevano essere anonimizzati).

4. Dati

Per poter applicare la procedura di anonimizzazione su sentenze con caratteristiche differenti, sono state prese in considerazione due fonti di dati (cfr. Tab. 3): le sentenze dalla Corte di Cassazione Penale sono emesse tra il 2011 ed il 2012 contenenti termini riferiti alla corruzione; le sentenze civili emesse dal Tribunale di Genova emesse tra il 2015 ed il 2019 derivanti dall'applicazione del Processo Telematico Civile.
I dati di input si differenziano non solo per l'oggetto della sentenza ma anche per le caratteristiche dei files di input: nel primo caso si è utilizzata la trasformazione in testo già disponibile sul sito, mentre nel secondo caso si è operata una trasformazione in testo dei files PDF.

Tabella 3. Corpus utilizzati[1]

Fonte	Giustizia	Tipologia	Num sentenze	Periodo	Num FG	Occorrenze
Corte di Cassazione	Penale	Corruzione	1.465	2011-2012	105.716	3.808.329
Tribunale di Genova	Civile	Divorzio	4.761	2015-2019	129.181	5.657.401

Il Corpus ottenuto dalle sentenze della Corte di Cassazione ha una maggiore variabilità della lunghezza delle sentenze, e mostra una maggiore ricchezza lessicale (cfr. Tabella 4), come era da attendersi. Le sentenze di divorzio del Tribunale di Genova hanno minori elementi descrittivi, specie in assenza di beni

[1] Le sentenze della Corte di Cassazione sono disponibili – in quanto ultimo grado di giudizio - sul sito SentenzeWeb della Corte di Cassazione; le sentenze di divorzio emesse nel periodo 2015-2019 sono disponibili grazie ad un accordo di ricerca tra la Scuola Superiore Sant'Anna ed il Tribunale di Genova.

materiali e/o di figli o nelle separazioni consensuali.

Tabella 4. Percentili delle occorrenze per oggetto delle sentenze

Quantile	Corruzione	Divorzio
100% Max	77.902	30.464
99%	20.551	5.047
75% Q3	2561	1.313
50% Mediana	1.592	907
25% Q1	1.004	694
1%	314	428
0% Min	182	345
Occorrenze/ FG	36,0	43,8
Media occorrenze per sentenza	2.599,5	1.188,3

5. Risultati

Sui due Corpus sono state applicate le procedure descritte nella Tabella 2, e successivamente è stata applicata la procedura di correzione dei due Corpus, generando due Corpus anonimizzati.

Su entrambi i Corpus sono state calcolate per ciascuna sentenza le variabili di conteggio degli elementi oscurati. Complessivamente sono state oscurate un numero non irrilevante di occorrenze (circa il 7,3% per i divorzi ed il 5,1% per la corruzione). Risultati più dettagliati sono riportati nelle Tabelle 5-7.

Tabella 5. Numero di elementi oscurati per fonte dei dati

	Corruzione			Divorzi				
	NOME	COGNOME	ENTEG	ANAG	NOME	COGNOME	CFISCALE	INDIRIZZO
Num occ oscurate	61.586	108.796	24.071	7800	117.318	237.317	14.660	12.661
% totale occorrenze	1,62	2,86	0,63	0,14	2,07	4,19	0,26	0,22

Il totale delle forme oscurate deriva soprattutto dal riconoscimento di nomi e cognomi e la percentuale media di forme oscurate per sentenza è pari a 7,1 (cfr. Tabella 6), con distribuzione simmetrica.

Tabella 6. Divorzi: distribuzione delle forme oscurate totali e loro percentuale per sentenza

	Min	P1	P10	P25	P50	P75	P	P99	Max	Mean	STD
Tot_oscurate	30	39	51	61	73	91	120	234	986	81,9	38,4
perc_oscurate	2,0	3,1	4,7	6,1	7,4	8,6	9,6	11,5	20,1	7,3	1,9

Tabella 7. Distribuzione delle sentenze per elementi oscurati per fonte dei dati

	Corruzione			Divorzi				
Statistiche	**NOME**	**COGNOME**	**ENTEG**	**NOME**	**COGNOME**	**CFISCALE**	**ANAG**	**INDIRIZZO**
min	4	4	0	5	5	0	0	0
P1	8	12	3	10	25	2	0	0
p10	16	29	7	15	15	2	2	2
Q1	30	56	14	19	19	3	2	3
mediana	81	132	31	27	27	4	4	4
Q3	244	444	68	40	40	5	4	6
P90	741	1.934	209	62	62	6	5	9
P99	1.861	3.966	650	152	191	10	8	23
max	1.861	3.966	650	405	405	20	12	51
media	42,0	74,3	16,4	24,6	49,8	3,1	1,6	2,7

Come era da attendersi esiste una fortissima correlazione tra numero di occorrenze e numero di nomi e cognomi oscurati (r=0,95), mentre la correlazione è più bassa (r=0,48) nel caso del riconoscimento delle persone giuridiche (non sempre presenti in tutte le sentenze).

Per quanto riguarda la differente distribuzione di NOME e COGNOME, il maggior numero di occorrenze di COGNOME deriva dalla presenza di riferimenti nella stessa sentenza al solo cognome dell'imputato o ricorrente oltre alla sua identificazione iniziale.

E' inoltre stato effettuato un controllo diretto attraverso la lettura di due campioni casuali di sentenze. L'individuazione delle sentenze costituenti i due campioni è avvenuta prima delle fasi di *text mining*: per le sentenze della Cassazione sono state selezionate 14 sentenze (estrazione con passo 100 dalla lista delle sentenze secondo l'ordine di input in TaLTaC) mentre per le sentenze civili ne sono state estratte 11 (estrazione con passo 400). Per ogni sentenza sono stati conteggiati i falsi negativi, cioè gli elementi non correttamente oscurati[2].

I risultati sono sintetizzati nella Tabella 8, dove riportiamo oltre ai risultati sull'identificazione complessiva di nomi e cognomi, anche quelli per i codici fiscali (presenti solo nelle sentenze civili provenienti dal Processo Telematico) e per gli Enti Giuridici (più rilevanti per le sentenze penali e quasi inesistenti per quelle civili).

Tabella 8. Principali risultati ottenuti dal controllo sulle 25 sentenze campionate

	Occorrenze totali nelle sentenze	**Totale di Nomi e Cognomi**	**di cui falsi negativi**		**Totale di Codici fiscali**	**di cui falsi negativi**		**Totale di Enti Giuridici**	**di cui falsi negativi**	
			n	**% sul totale**		**n**	**% sul totale**		**n**	**% sul totale**
Corruzione	28.769	1.038	*44*	4,2	--	--	--	84	*34*	40,5
Divorzi	11.038	649	*26*	4,0	44	*2*	4,5	--	--	--

Per maggiore sicurezza è stato inoltre effettuato un ulteriore campionamento di 200 sulle 944 sentenze civili per le quali non veniva riconosciuto nessun codice

[2] Per brevità e per rispetto della normativa della privacy non viene qui riportato l'elenco delle sentenze sottoposte a lettura di controllo, con il dettaglio del numero e tipo di elementi non correttamente oscurati (falsi negativi). I dati sono comunque disponibili a richiesta.

fiscale: una lettura limitata alla sola parte introduttiva della sentenza ha evidenziato solo 10 casi in cui non è avvenuto il riconoscimento, mentre per le restanti 190 non è stato trovato nessun codice fiscale nella scrittura della sentenza.

6. Conclusioni

I valori ottenuti sono complessivamente molto positivi: in molte sentenze il numero dei falsi positivi per i Nomi e Cognomi è pari a 0 (4 sentenze di corruzione e 5 di divorzio), ed i falsi negativi derivano dal mancato riconoscimento di alcuni cognomi. Il riconoscimento dei Nomi è completo nelle sentenze sottoposte a lettura, grazie alle risorse linguistiche di TaLTaC. Inoltre, appare ottimale anche il riconoscimento dei nomi e cognomi di magistrati e di avvocati.

Maggiori problemi pone il riconoscimento degli Enti Giuridici, sia per le complesse modalità di scrittura della forma giuridica, sia per errori dovuti alla conversione in testo dei files originari.

I codici fiscali sono riconosciuti correttamente in tutte le sentenze di divorzio eccetto una sola, derivante da una scrittura errata di uno stesso codice fiscale ripetuto due volte nella stessa sentenza. La categoria semantica CFISCALE è stata assegnata anche alle non poche forme grafiche scritte in modo non corretto, pari al 19,6% dei codici fiscali identificati.

Bibliografia

Bolasco S. (2010). *TaLTaC 2.10: sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. LED, Milano.

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di statistica,* vol 7: 17–53.

Bolasco S., D'Avino E. and Pavone P. (2007). Analisi dei diari giornalieri con strumenti di statistica testuale e text mining. In: *I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo*. ISTAT, pp. 309–340.

Bolasco S. and De Gasperis G. (2017). TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences, in: *Data Science and Social Research*. Springer, pp. 97–103.

Csányi G.M., Nagy D., Vági R., Vadász J.P. and Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. *Symmetry* vol 13 (8): 1490.

Glaser I., Schamberger T. and Matthes F. (2021). Anonymization of german legal court rulings, in: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. pp. 205–209.

Ohm P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *57 UCLA LAW REVIEW*: pp. 1701-1777.

Povlsen C., Jongejan B., Hansen D.H. and Simonsen B.K. (2016). Anonymization of court orders, in: *Iberian Conference on Information Systems and Technologies (CISTI).* pp. 1–4. doi: 10.1109/CISTI.2016.7521611

Romano M.F., Baldassarini A. and Pavone P (2020). Text Mining of Public Administration documents: preliminary results on judgements, in: *Text Analytics*, Springer, pp. 117-126.

An analysis of the Jesus' community in the Gospel of Matthew, Mark, Luke, John, and Acts of Apostles networks[1]

Roberto Rondinelli[1], Stefano Marmani[2], Valerio Ficcadenti[2]

[1]University of Naples Federico II – roberto.rondinelli@unina.it
[2]London South Bank University – marmanis@lsbu.ac.uk
ficcadv2@lsbu.ac.uk

Abstract
The Bible has always attracted the interest of scholars from different fields, specifically, Jesus' paths taken out from the writings are at the centre of an ongoing discussion in the literature. In this short paper, we study the relationship between Jesus and other biblical characters and places via a text mining network-based approach. From the World English Bible, we take the Gospel of Matthew, Mark, Luke, John, and Acts of the Apostles and analyse the resulting corpus. Namely, people and places' names are collected using a POStagger on the corpus and manually checked for finalising the pre-processing. Then, the selected names' co-occurrences are searched in each verse. It results in an adjacency matrix containing the frequencies of names' co-occurrences and a set of nodes' attributes. The strength of links between a couple of names is defined as the sum of times that the couple appears in all verses. Such a network is used to explore communities of names around Jesus. To do so, we use the Louvain community detection method for each sub-network referring to every book.

Keywords: Text Mining, Networks, Gospels, Acts of Apostles, Jesus' communities

1. Introduction

Addressing texts' frameworks via word-based Networks is becoming more and more common. When such a method is used, words are nodes, and co-occurrences in the texts are their edges (see Lane and Gantley, 2017). In other contexts where natural language processing techniques are used, texts are intended as nodes and the number of shared words as their links. In this paper, we employ a text mining network-based approach to study the relationship between names of people and places in the Gospel of Matthew, Mark, Luke, John

[1] We thank Roberta Baglioni for her input in the interpretation of the results.

and Acts of the Apostles from the World English Bible (WEB hereafter, see Wikipedia contributors, 2019), with a specific focus on the Messiah and its neighbourhood framework. The idea of studying the characters and places' relationships via a text mining network-based approach is used to verify some well-known patterns in Biblical history, providing a socio-spatial view. In doing so, we move our steps from Rondinelli et al. (2020a,b) improving the quality of the data and enhancing the analysis of the Jesus' community. In fact, we have a new list of proper names generated utilising the NLTK's part-of-speech tagger (Bird et al., 2009) on the WEB, and then we manually assess the resulting terms to exclude flaws in the POStagging step (details on this are presented in the next session).

The study of the Bible brings together numerous disciplines, making it an ideal setting for inter-disciplinary research. It is worth noting the examples from Mehri and Jamaati (2021); Honoré (1968), and Agata and Agata (2021). In the context of this research, we address the narrative routes of the five biblical books mentioned above via the presence of the proper names and their in-verse relationships. The analytical tools are drawn from the discipline of text mining and the methods from network theory. This technique is gaining popularity due to the opportunities it provides. Cinelli et al. (2019) provide an example of information gained from a similar technique to studying United States of America presidents' statements. Rochat (2014) investigates the principal personalities of Les Confessions (Geneva version of the manuscript), and later Rochat (2015) gives an outstanding analysis of the character framework in Emile Zola's twenty books that comprise Les Rougon-Macquart, published between 1871 and 1893. Referring to biblical studies, from a network theory point of view, we are in line with Czachesz (2016) and Zhitomirsky-Geffet and Prebor (2019), but we go further with the Community Detection analysis getting close to Lane and Gantley (2017) but providing a focus on Jesus. From a text mining point of view, we differ from Czachesz (2016) because we do not build a semantic network as the author did. Therefore, we do not deal with issues like fully connected networks of words in sentences or with the idea of using almost all the words (semantic approach) in verses; we use the list of names resulting in the "NNP" class from the application of the POStagger and refined manually by us (see Bird et al., 2009, for the details on the POStag's classes). We use the verse as the smallest unit of analysis, as recommended by Bar-Efrat (1980) and in line with Beveridge and Shan (2016), who studied Game of Thrones books with a very similar approach to the one we are using. For a wider view on the literature, see Rondinelli et al. (2020a,b). In Section 2, we present the data improvements, the network formalisation, and the method for the analysis. Section 3 contains the results and the main conclusions.

2. Data, Network and Method

The Bible utilised in this study is the World English Bible (Wikipedia contributors, 2019). It was obtained from Hartono (2019). The website delivers the 66 Bible books (Old and New Testament) already versed in SQL format. Our study focuses on the four Gospels (Matthew, Mark, Luke, and John) and the Acts of the Apostles, which are 5 of the 27 New Testament books.

To reach our research objectives, we need a list of proper biblical names divided into names of people and names of places. Such a list is obtained by deploying the steps presented below:

1. To fetch the 5 books from Hartono (2019). They are already organised in SQL format.
2. To remove footnotes from the raw texts.
3. To remove punctuation.
4. To tokenise the verses.
5. To apply the NLTK's POStagger (Bird et al., 2009) and select the tokens classified as "NNP" (proper noun) from each verse.
6. To manually check for flaws in the POStagging and classify the names in order to have dichotomous classes: names of people and names of places. For example, we removed "Baptised" or "Blood"; the POStagger captures them as "NNP", but they are not proper names of people or places.

The resulting list of proper names is used to search for their presence in each verse and register their intra-verse co-occurrences. We build an overall network of all considered books (corpus) that can be formalised as a one-mode attributed and weighted words network representing the relationships among proper names of the Gospel of Matthew, Mark, Luke, John, and Acts of Apostles. This network G consists of a 4-tuple $G\ (V, R, W, A)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of n proper names; $R \subseteq V \times V$ is the set of the edges $r_{ij} = (v_i, v_j)$, $r_{ij} \subseteq R$ where $i = 1, \ldots, n$ and $j = 1, \ldots, n$; i.e. each edge is a couple of nodes (proper names) and indicates whether v_i occurs with v_j at least in a verse; $W : R \longrightarrow \mathbb{N}$, with $W(r_{ij}) = w_{ij} \subseteq \mathbb{N}$ representing the number of times v_i occurs with another proper name v_j in the verses of the considered corpus. The nodes are featured by the set of attributes $A = \{A_1, A_2, \ldots, A_K\}$, where $k = 1, \ldots, K$, and $A_k : V \longrightarrow dom(A_k)$, with $A_k(v_i) = a_{ik} \subseteq dom(A_k)$ the generic attribute of the proper name v_i, with $dom(A_k)$ the set of values related to a generic attribute. In this case, we have $K = 6$ dichotomous attributes, five for books, each representing whether a proper name is observed in a specific book, and one for determining if the name belongs to a person or a place. By means of these attributes, we can study the sub-networks referring to specific books or specific types of proper names.

By employing the five attributes related to the considered books and thus selecting their five sub-networks, we look for the community of Jesus. Such a

community is seen as the circle of people and places' names strictly connected to the Messiah's one (and his neighbours) and scarcely connected to the rest of the network. We apply the Louvain method (see Blondel et al., 2008) to pursue this purpose, and we use the new and refined list of names which is an improvement with respect to Rondinelli et al. (2020a,b).

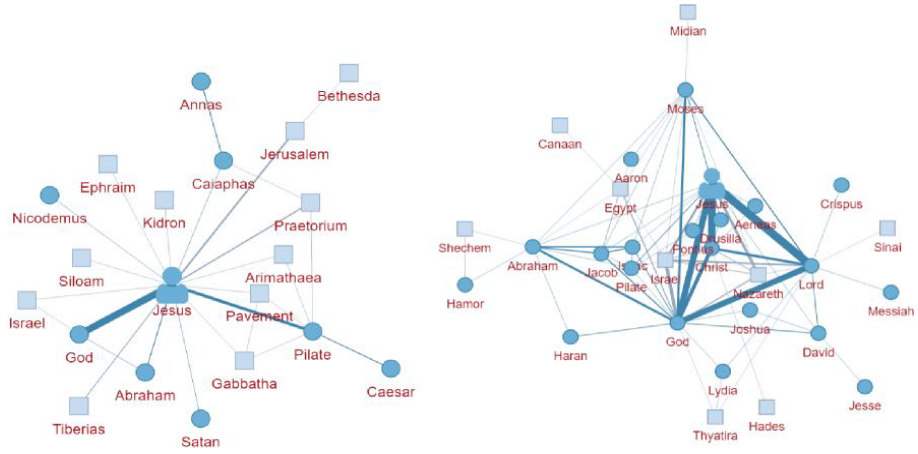
3. Results and Conclusions

The community detection analysis divides each network into groups of nodes densely connected within them-self and poorly connected with the rest of the network. Given the resulting communities, we focus on those containing Jesus, which are reported in Figures 1 and 2. Figure 1 contains the Jesus' communities generated in the Gospel of John and Act of Apostles, while Figure 2 reports the ones of the Synoptic Gospels (Mark, Matthew, and Luke). Interestingly, some widely studied aspects emerge for the Jesus' communities in the Synoptic Gospels (see Goodacre, 2004, for an overview of salient themes regarding the comparison of Mark, Matthew, and Luke's Gospels). Figure 2 shows that the connection between Jesus and God is present in the 3 communities, and its intensity is similar to the strongest one appearing in Mark's Gospel. The triad God-Jesus-Lord is also present in all the Synoptic books, but Christ-God-Jesus is absent in Luke because Christ is not connected to Jesus. Satan appears only in Matthew and Luke, being part of two relevant triads: God-Jesus-Satan and Jesus-Lord-Satan. In Luke, Satan also connects the triad Iscariot-Judas-Satan. Regarding places, Israel plays different roles. In Mark, it is connected to Christ, God, Jesus, and Lord, but in Matthew, it is connected only to Jesus, while in Luke, it is connected with God, Jesus and Lord. Jerusalem has a role in Luke, connecting Jesus, God and Lord with Cleopas, Siloam and Emmaus acting as a bridge. In carrying on with this exercise of comparing the presence and absence of relationships, one can determine the differences in the stories where Jesus is at the centre of the communities, tracing back known hypotheses in the synoptic problem's literature, like those in regard to the *Q-source* (see, Goodacre, 2004). Another important aspect that emerges from Jesus' communities in Figure 2 is the episode of his death, from Pilate to his burial in Arimathaea. This is evident in the Gospel of Mark and Matthew, where a motif of four nodes composed by Jesus-Pilate-Barabbas-Arimathaea is observed with two different configurations.

On the other hand, the peculiar case of the Gospel of John (Figure 1) presents a star-like Jesus' community shape with a few triads mostly related to places, indicating a more geographical narrative than the one in the Synoptic Gospels. In the Gospel of John, we also distinctly recognise the episode of the Jesus' death through the sub-group Jesus-Pilate-Pavement-Arimathaea-Gabbatha, which in Act of Apostles appears only thanks to the presence of Pilate and Pontius (please

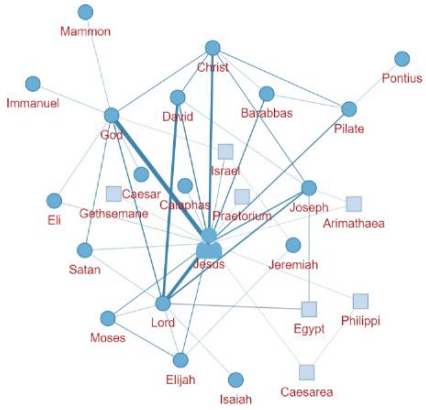
note that in our analysis, the full name Pontius Pilate results in two different names, so in the network they are represented by 2 nodes). In the Act of Apostles, one can notice a remarkable structure of simplicial complex type creating a meaningful and robust structure between Christ, God, Jesus and Lord, all strongly related to Israel. A visual inspection of the Act of Apostles presents similarities with Luke's Gospel in terms of network structure.

In this case, the studied Jesus' communities offer a good perspective on events closely related to the Messiah and assert the goodness of the methodology. In fact, Figures 1 and 2 confirm and present important passages of the Bible's plot.

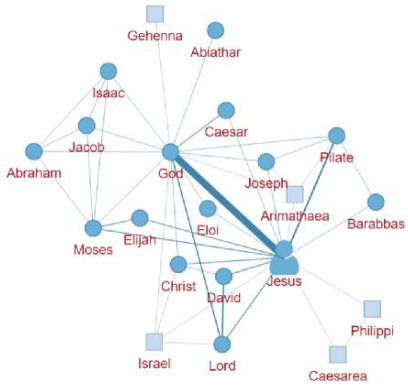


(a) Gospel of John *(b) Acts of Apostles*

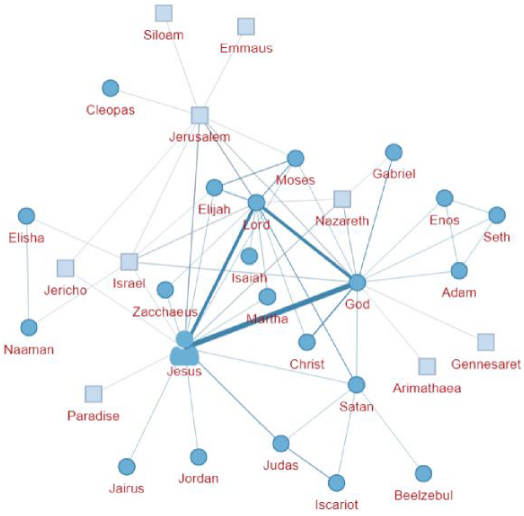
Figure 1. The networks of Jesus' communities from the Gospel of John and Acts of Apostles. The circled dots are referred to people and the squared ones are used for places.



(a) Gospel of Matthew



(b) Gospel of Mark



(c) Gospel of Luke

Figure 2. The networks of Jesus' communities from the Gospel of Matthew, Mark, and Luke. The circled dots are referred to people, and the squared ones are used for places.

References

Agata M. and Agata T. (2021). Statistical analysis of the Gutenberg 42-line Bible Types: Special Focus on Letters with a Suspension Stroke. *The Papers of the Bibliographical Society of America,* 115(2): 167-183.

Bar-Efrat S. (1980). Some observations on the analysis of structure in biblical narrative. *Vetus Testamentum*, 30(2): 154-173.

Beveridge A. and Shan J. (2016). Network of thrones. *Math Horizons*, 23(4): 18-22.

Bird S., Klein E. and Loper E. (2009). *Natural language processing with Python: analysing text with the natural language toolkit*. O'Reilly Media, Inc.

Blondel V. D., Guillaume J. L., Lambiotte R. and Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

Cinelli M., Ficcadenti V. and Riccioni J. (2019). The interconnectedness of the economic content in the speeches of the us presidents. *Annals of Operations Research*, 299(1): 593-615.

Czachesz I. (2016). Network analysis of biblical texts. *Journal of Cognitive Historiography*, 3(1-2): 43-67.

Goodacre M. (2004). *The synoptic problem: A way through the maze*. Vol. 80. A&C Black.

Hartono O. R. (2019). Bible corpus. https://www.kaggle.com/oswinrh/bible. [Online; accessed 5-February-2020].

Honoré A. M. (1968). A statistical study of the synoptic problem. *Novum Testamentum*, 10(Fasc. 2/3): 95-147.

Lane J. E. and Gantley M. J. (2017). Utilising Complex Systems Statistics for Historical and Archaeological Data. *Journal of Cognitive Historiography*, 3(1-2).

Mehri A. and Jamaati M. (2021). Statistical metrics for languages classification: A case study of the Bible translations. *Chaos, Solitons & Fractals*, 144: 110679.

Rochat Y. (2014). Character Networks and Centrality. *PhD thesis*, University of Lausanne.

Rochat Y. (2015). Character network analysis of Emile Zola's Les Rougon-Macquart. *Proc. of DH 2015 (Digital Humanities 2015), Sydney.*

Rondinelli R., Marmani S. and Ficcadenti V. (2020a). Biblical names' relationships in the gospel of matthew, mark, luke, john and acts of apostles. *arXiv preprint arXiv:2012.04753*.

Rondinelli R., Marmani S. and Ficcadenti V. (2020b). Biblical names' relationships in the gospel of Matthew, Mark, Luke, John and Acts of Apostles. *Lexicometrica*. http://lexicometrica.univparis3.fr/jadt/JADT2020/.

Wikipedia contributors (2019). World english bible — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=World_English_Bible&oldid=903093909. [Online; accessed 6-February-2020].

Subjectivity and connectors: study of the parameters of discourse genres and languages[1]

Corinne Rossari, Cyrielle Montrichard, Claudia Ricci, Linda Sanvido

Université de Neuchâtel – corinne.rossari@unine.ch
cyrielle.montrichard@unine.ch claudia.ricci@unine.ch
linda.sanvido@unine.ch

Abstract
In this article we use two statistical approaches to study the functioning of different categories of connectors in relation to the class of adverbs. We analyse their behaviour in corpora representing different genres (press and encyclopaedia), in French and in Italian. The first method mobilizes co-occurrence specificity scores to identify the collocation profile of each connector, permitting us to class them on a scale depending on their subjectivity charge. The second method uses factorial correspondence analysis (CA) to graphically represent groupings of connectors. The CAs can be in line with the subjectivity scales, but they can also make appear characteristics that the latter do not show, thus supporting the interest of two distinct but complementary approaches.

Keywords: connectors, adverbs, subjectivity markers, co-occurrence, factorial correspondence analysis

1. Introduction

Our contribution proposes a textometric approach which is innovative on an analytical level and experimental on a methodological level: we aim to evaluate the behavior of a functional class of words, connectors. For this purpose, we will proceed to the identification of their co-occurrents, in different informative text genres in two Romance languages, French and Italian. The interest of this class of words lies not only in its transversal nature – compared to lexical categories, connectors are less dependent on constraints linked to themes or discourse genres – but also in the fact that they shed light on the text structure on an argumentative level.

[1] This article is a contribution to the research project The argumentative patterns of modal forms. A corpus-based study comparing French and Italian (n. 100001F_192247) funded by SNF.

We compare corpora representing two genres of informative discourse in both languages: regional press (*L'Est Républicain, L'Adige*) and encyclopedic discourse (French and Italian *Wikipedia*)[2]. Our research focuses on a paradigm of connectors representing different classes of primitive cognitive relations: cause, purpose, opposition and temporality. To identify the behavior of connectors in these corpora we consider the way in which they combine with another word class: the adverbs in *-ment/-mente*, since they also play an important role in the argumentative construction of a text, insofar as most of them involve the speaker's voice. Our methodology is based on the crossing of two statistical methods[3]: (i) co-occurrence specificity scores (cf. Lafon, 1980) between connectors and adverbs to identify the collocation profile of each connector; (ii) graphical representations using co-occurrence factorial Correspondence Analysis (CA) – cf. Mayaffre, 2014; Viprey, 2016 – to identify re-/un-groupings between connectors according to their co-frequency with the thirty most frequent adverbs of each corpus.

The first part of our article presents the two word classes involved in our quantitative analysis: the paradigm of French and Italian connectors and the classes of adverbs. The second part exposes the analysis of the data extracted thanks to the co-occurrence scores and the third part the analysis of the graphs of the CAs.

2. Paradigm of French and Italian connectors and the class of adverbs selected

Connectors are usually considered as playing a major role in the argumentative construction of a discourse (cf. Ducrot et al., 1980; Anscombre, 2002; Roulet et al., 1987). We therefore assume that their use is more or less sensitive to the subjectivity of the speaker: an opposition with *mais* makes appear, more strongly than with *pourtant*, the presence of the speaker as the person responsible for conceiving the opposition. This is the reason why *mais* allows relations between any property whereas *pourtant* does not (it only works when an opposition is easily retrievable from the propositional content).

Tous les étudiants ont réussi, mais/ ?pourtant l'examen était facile

[2] *L'Est Républicain* (ER), 25'825'949 tokens, all articles published between 2010.01.02 and 2010.02.26 and between 2010.03.17 and 2010.04.30; *L'Adige* (AD), 20'343'148 tokens, 38'496 articles covering the period 1996-2006; French *Wikipedia* (FR-WIKI), 18'715'455 tokens, 10'571 articles, from 2019 (without bibliographic sections); Italian *Wikipedia* (IT-WIKI), 18'394'987 tokens, 15'905 articles from 2019 (without bibliographic sections). All articles have been randomly selected.

[3] All queries and statistical calculations have been done with TXM (Heiden et al., 2010) and Hyperbase to build CAs (Brunet, 2012).

('all students passed, but/yet the exam was easy')[4]
We can observe similar differences with causal connectors:
Je trouve que c'est un examen difficile, donc/ ?de ce fait Paul est intelligent
(I find this exam difficult, so/as a result, Paul is smart)
Donc allows causal connections involving personal judgments, whereas *de ce fait* seems less akin to express such links.
To assess their subjectivity load, we will look at the combinatorial properties of connectors with a class of words which also involve the speaker's subjectivity: adverbs in *-ment*. Among these, sentence adverbs of enunciation and modality – categorized in an inventory by Molinier and Levrier (2000) as "adverbes disjonctifs de style" (enunciation) and "adverbes modaux" (modality) – are specialized in conveying a speaker's voice commenting her own enunciative act (such as *franchement*, 'frankly') or her judgment on the event (*réellement*, 'really'). Such a comment materializes the presence of the speaker in her own speech, especially compared to other classes of adverbs, for instance manner adverbs (*lentement* 'slowly'). This is why we consider that these two categories are particularly relevant to assess the presence of the speaker.
Our list represents the main cognitive categories. We have selected some entities for each category in French paying attention to exclude the forms that present homonymy cases. From this list, the Italian connectors have been selected by consulting Europarl's parallel corpus as well as Reverso.com as a reference for translation in context. A further selection has been made according to the specificities of connectors. We have retained *mentre* as the Italian counterpart of *alors que* and *tandis que* and not *quando*, because the high frequency of the contrastive value of *mentre* makes it closer to *alors que/ tandis que* than *quando*. *Perché* has been excluded as correspondent of *car* because, it also translates *pourquoi* ('why') and *pour que* ('in order to').

Table 1. Paradigm of selected connectors in French and Italian

	French	Italian
consequence	donc, de ce fait ('so', 'therefore')	dunque, quindi
cause	car, puisque ('because', 'since')	poiché, visto che
purpose	afin que, pour que ('in order to', 'so that')	affinché
opposition	mais, cependant, pourtant ('but', 'however', 'yet')	ma, però, eppure
contrast	alors que, tandis que ('whereas')	mentre
temporality	puis, ensuite ('then', 'after that')	dopo, poi

[4] The English translations are given for readers' convenience, but are not part of the analysis.

3. Co-occurrence specificity scores between connectors and adverbs

We searched for all of the adverbs in *-ment/-mente* that are specific co-occurrents[5] of each of the selected connectors, and ranked the connectors on a scale (see below) going from 'less subjective' to 'more subjective' according to the number of specific adverbs each one has[6]. For the reason given above, we have counted twice the enunciation and modal adverbs[7]. The first number following each connector corresponds to the total of its specific adverbs, the second to the total including adverbs counted twice. Using the co-occurrence specificity scores – which are not impacted by raw frequencies – allows us to question any connectors and adverbs provided that there was a minimum frequency and co-frequency of 2.

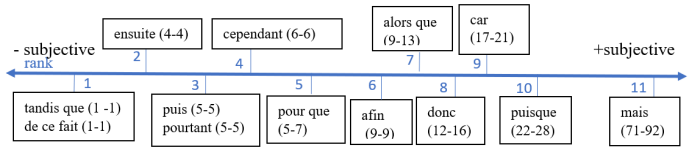


Figure 1 - L'Est Republicain : subjectivity scale

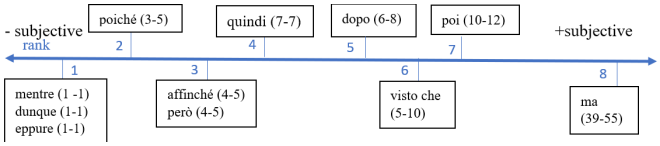


Figure 1. L'Adige : subjectivity scale

The most striking results these scales provide is that *mais* and *ma* are the connectors receiving the highest number of adverbs including the enunciative and modal adverbs, and this makes them appear as the most subjective connectors. On the other hand, *donc* and *dunque* occupy opposite positions: *donc* is close to the pole indicating a higher degree of subjectivity, *dunque* is close to the opposite pole. This allows us to notice that connectors conveying the same

[5] Since the relation between the connector and the adverb is not annotated, we tested multiple query spans. We choose ten items before the connector and five after it, in order to maximize the chances of collecting adverbs in first position within the sentence and to minimize the chances that the adverb after the connector is not relevant.

[6] A representation based on statistical significance of joint co-occurrences would allow us to see the links between connectors in relation to their common specific adverbs. However, we would lose the ranking representation, which is relevant for the connectors' semantic proprieties as explained in section 2.

[7] The reader can find all the statistical details about specific adverbs for each connector in: https://drive.switch.ch/index.php/s/P6s9iQy4hRZU702.

type of relation and being morphologically equivalent (*donc/dunque*, *mais/ma*) can differ strongly or, on the contrary, be completely homogenous in what concerns their subjectivity charge. In addition, we observe that the connectors conveying a relation not between propositional contents but between two speech acts – *car* and *puisque* (cf. Groupe Lambda-l, 1975) are close to the 'more subjective' pole. Such a correlation shows that the level of the relation conveyed by the connectors seems to play a major role in their capacity to attract adverbs. By looking at a corpus representing another genre we can measure the stability or not of the subjectivity load of a connector.

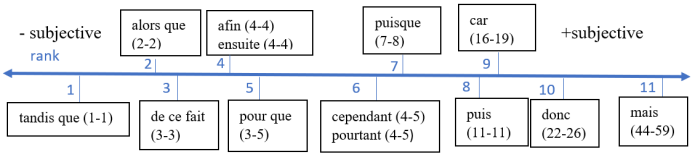


Figure 2. French Wikipedia: subjectivity scale

Many connectors behave similarly in both corpora: *mais* occupies the highest place on the scale; *donc* is also close to the higher subjectivity pole, whereas *de ce fait* to the opposite one. Some differences concern *alors que* and *puis*: the former is closer to the 'less subjective' pole in FR-WIKI compared to ER, whereas *puis* is closer to the higher subjectivity pole in FR-WIKI compared to ER. In terms of stability, these data show that *mais, donc* and *de ce fait* are stable, whereas *puis* or *alors que* are more subject to genre-related differences.

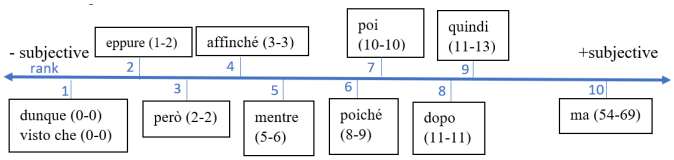


Figure 3. Italian Wikipedia: subjectivity scale

We observe the same tendencies in Italian: *ma* is located on the 'more subjective' pole in both corpora; *quindi* and *dunque* show the same behavior: they are located at the extremities of the scale (more subjective for *quindi* and less for *dunque*). As in French, differences between the two corpora appear, showing also some instabilities: *visto che* is located on the 'less subjective' pole in IT-WIKI and near the 'more subjective' pole in AD.

The ranking above has allowed us to establish (i) similarities between French and Italian connectors, (ii) stabilities and variations in the subjectivity load

according to the corpus; however, it has not given any indication concerning the global behavior of the connectors in relation to the class of adverbs. We cannot have an idea of the role of the cognitive relation they convey or of their subjectivity load on their behavior with adverbs. In other words, do connectors tend to form clusters with connectors sharing the same kind of relations, or the same kind of subjectivity load, or both, or neither? CA method allow us to challenge these questions.

4. Correspondence Analysis on the matrix Adverbs x Connectors

The CAs[8] are built extracting the thirty most frequent *-ment/-mente* adverbs of each corpus, exploiting a set of data that is different from the one used in the ranking method. They offer a view of the proximity and distance between connectors themselves as well as between connectors and adverbs.

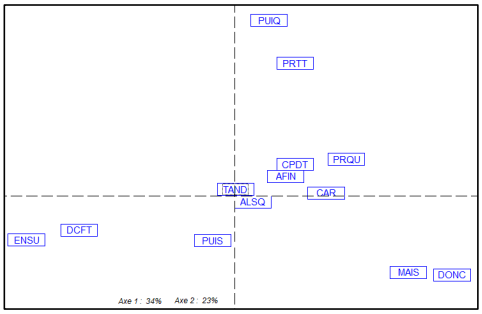


Figure 4. CA of columns headers in FR-WIKI

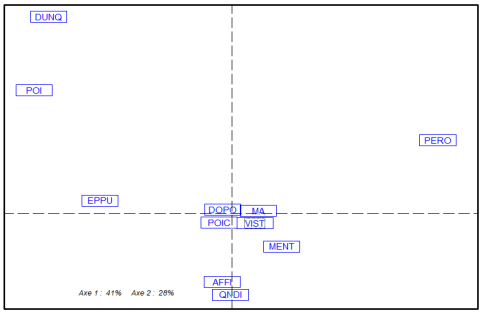


Figure 5. CA of columns headers in IT-WIKI

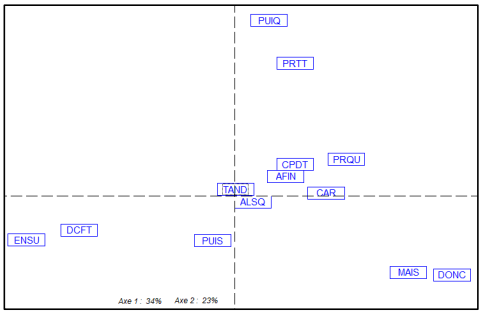


Figure 6. CA of column headers in ER

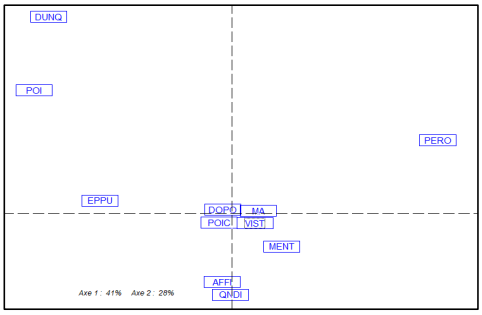


Figure 7. CA of column headers in AD

[4] The matrix is based on the co-frequency between the adverbs (line headers) and tested connectors (column headers), using the same span as for the co-occurrence specificity scores. For space reasons, only column headers are presented but complete CAs are available here: https://drive.switch.ch/index.php/s/P6s9iQy4hRZU702

First of all, looking at the off-center connectors, in both French corpora, donc and mais form a cluster, showing that beyond their proximity on the subjectivity scale they also combine similarly with the most frequent adverbs. This regularity in their behavior is observable even if we look more specifically at the adverbs appearing around them: in both CAs we find seulement, 'only'; finalement, 'ultimately'; également, 'also'. Such regular results from the two different methods and dataset could be surprising especially since the selected adverbs for the two corpora do not show the same balance between non-enunciative and enunciative adverbs (8/30 enunciative adverbs in ER and 4/30 in FR-WIKI, all different). This shows that, more than their relational properties, it is their subjectivity load that guide the way they combine with ment- adverbs. More generally, the CAs of both corpora show that the combination of off-center connectors with adverbs crosses the boundaries of the semantic relation they express: in addition to the proximity between mais and donc (which represent two cognitive categories) we observe a cluster between de ce fait and ensuite, puisque and pourtant in FR-WIKI and between pourtant and puis in ER. These clusters are consistent with the subjectivity scale, but we can also point out a grouping that is not: ensuite and puisque in ER. Otherwise, the CAs show for both corpora a grouping around the middle of the axes concerning connectors that are semantically very closely related, namely alors que and tandis que, pour que and afin que, which can have different positions on the subjectivity scale – cf. tandis que and alors que (respectively rank 1 and 7 in ER). Their semantic proximity seems to take precedence over their subjectivity charge in this corpus, unless, since they are near the center of the graph, this proximity should be interpreted as a result of the banality of their combinatory properties.

The CAs concerning the two Italian corpora show a data flattening on the middle-lower part of the graph: *ma* unexpectedly comes in the middle of the graph in IT-WIKI and slightly off-center in AD. This is related to the peculiar spot of some connectors at the extremities of the upper quadrants, due to their very high contribution to the building of both axes. Therefore, the graph needs to be analyzed looking at (i) *quindi* and *eppure* in AD that are isolated at each extremity of axis 1; (ii) *dunque* and *poi* in IT-WIKI, that form an isolated cluster in the upper-left quadrant. Such a position for *quindi* and *eppure* does match the one on the subjectivity scale in AD but for *dunque* and *poi* in IT-WIKI it is not the case (they are positioned far from each other on the scale). These discrepancies in IT-WIKI can result from the fact that enunciative adverbs are particularly rare in the first thirty more frequent adverbs of this corpus (only two items).

5. Conclusion

Crossing two statistical methods using different datasets has allowed to outline a similar behavior between connectors pertaining to different semantic classes as well as a different behavior between morphologically related connectors in French and Italian. It also challenged the factor of genre on the general behavior of the selected connectors. It emerges that the latter has an impact, but weak, since more obviously in French than in Italian the results issued in both corpora are in line, be it with co-occurrence specificity scores or with CAs. In Italian, the difference concerning the results issued from CAs should be studied in more depth as graphs are built only on some connectors that absorb a high percentage of the information.

References

Anscombre J.-C. (2002). Mais/pourtant dans la contre-argumentation directe : raisonnement, généricité, et lexique. *Linx*, 46: 115-131.

Brunet E. (2012). Nouveau traitement des cooccurrences dans Hyperbase. *Corpus* [en ligne], 11. URL : http:// journals.openedition.org/corpus/2275

Ducrot O. et al. (1980). *Les mots du discours*. Paris: Minuit.

Groupe Lambda-l (1975). Car, parce que, puisque. *Revue Romane,* 10: 248-280.

Heiden S, Magué J.-P. and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010*. Rome, Italie. URL: http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). *JADT 2014*. Paris, France, pp.15-32.

Molinier C. and Levrier F. (2000). *Grammaire des adverbes. Description des formes en -ment*. Genève/Paris : Droz.

Roulet E., Auchlin A., Moeschler J. and Rubattel C. (1987). *L'articulation du discours en français contemporain* (Vol. 11). Peter Lang Gmbh, Internationaler Verlag Der Wissenschaften.

Viprey J.-M. (2016). Comparer des AFC de cooccurrence généralisée. *JADT 2016*. Nice, France. URL: http://lexicometrica.univ-paris3.fr/jadt/jadt2016/01-ACTES/83505/83505.pdf

Les marqueurs discursifs : étude comparative de leur utilisation et fréquence chez les locuteurs francophones et dans les manuels de français langue étrangère

Simona Ruggia, Minerva Rojas

[1]Université Côte d'Azur – Simona.RUGGIA@univ-cotedazur.fr
Minerva.ROJAS@univ-cotedazur.fr

Abstract
This study examines the use of several discourse markers (DMs) in oral interactions between French native speakers and those used in textbooks for teaching French as a foreign language (FFL). Two corpora (corpus Manuels FLE and corpus SCFLE) are examined to explore possible differences in the frequency of occurrence of a selection of DMs and to calculate a rate of marker use. The results show that the distribution of DMs differs between the two corpora and that their proportion is higher in spontaneous French than in textbook interactions.

Keywords: discourse markers, oral interactions, French native speakers, French as a foreign language textbooks.

Résumé
Cette étude se penche sur l'utilisation de certains marqueurs discursifs (MD) lors d'interactions orales entre locuteurs natifs de langue française et de celles exploitées dans les manuels destinés à enseigner le français langue étrangère (FLE). À travers l'analyse de deux corpus (corpus Manuels FLE et corpus SCFLE), nous explorons les éventuelles différences en repérant la fréquence d'occurrence d'une série de MD et en calculant leur taux d'utilisation. Les résultats montrent que la distribution des MD diffère entre les deux corpus et que leur proportion est plus élevée chez les locuteurs natifs.

Mots clés : marqueurs discursifs, interactions orales, locuteurs francophones, manuels de français langue étrangère.

1. Introduction

Les marqueurs discursifs (MD) du français parlé ont suscité un intérêt remarquable à partir des années 1970 en donnant lieu à de nombreuses descriptions, classifications et terminologies issues de diverses approches linguistiques. Il convient donc de préciser que nous utiliserons le terme de MD

« dans un sens large, en tant que classe hétérogène » (Rodriguez Somolinos, 2011 : 6) comprenant des unités lexicales normalement utilisées à l'oral (Dostie et Pusch, 2007), qui ne sont pas nécessairement des unités référentielles car leur absence ne suppose pas une perte d'information (Chanet, 2004) mais qui fournissent des pistes à l'interlocuteur sur la façon dont le locuteur construit son discours et connecte les contenus (Bazzanella, 2006).

Dans le cadre de notre étude, nous nous intéressons aux MD d'un point de vue linguistico-didactique, afin de comparer leur utilisation et fréquence dans deux corpus numériques composés d'interactions orales de nature différente. D'une part, un corpus constitué d'interactions (dialogues et polylogues) proposées sous forme d'enregistrements et de textes dans plusieurs manuels de français langue étrangère (FLE) couvrant tous les niveaux[1] du *Cadre Européen Commun de Référence pour les langues* (CECRL) (Conseil de l'Europe, 2001). D'autre part, un corpus comprenant des interactions (dialogues) de locuteurs francophones.

Pour ce faire, nous allons nous pencher sur l'emploi des MD à travers l'analyse quantitative et qualitative de deux corpus, afin de vérifier si les modèles présentés dans les manuels de FLE reproduisent les caractéristiques d'une interaction orale spontanée ; si les MD les plus utilisés dans les manuels sont ou non les plus employés dans les interactions spontanées ; et si la fréquence d'utilisation des MD est significativement différente dans les manuels par rapport au français parlé spontané.

2. Cadre théorique

Afin de décrire les emplois de certains MD présents dans nos deux corpus, nous nous appuierons sur la taxonomie esquissée par Waltereit (2007 : 95). Ainsi, nous allons distinguer les MD qui « existent sous la forme d'un seul mot » (*Ib*. : 94) comme *mais, donc, alors*, etc. que nous appellerons simples en les opposant aux marqueurs « complexes » qui sont constitués de « deux [ou plusieurs] mots contigus » (*Ib*. : 94). Ces derniers correspondent aussi bien aux MD phrasèmes (Dostie, 2004) constitués de mots qui ne sont pas des marqueurs comme *tu sais, tu vois, tu sais pas quoi*, qu'aux MD composés d'autres MD comme *bon ben, enfin bref, et puis ben,* etc. N'ayant pas la place ici pour détailler le rôle des MD lors des interactions verbales, nous nous contenterons d'affirmer qu'ils fournissent un cadre interprétatif dans lequel le message doit être compris.

3. Méthodologie

L'exploration et l'analyse des corpus a été effectuée grâce à une lecture contrôlée et assistée par l'analyse statistique des données textuelles, la logométrie

[1] Le CECRL propose une classification de la compétence à communiquer langagièrement en 6 niveaux : A1, A2, B1, B2, C1 et C2. (Conseil de l'Europe, 2001: 23-38).

(Mayaffre, 2010) qui a « une dimension métrique et réflexive par rapport aux observables du texte » et « un pouvoir descriptif qui a une valeur quasi probatoire » (Ruggia, 2021 : 88).

Les corpus ont été constitués et étudiés à l'aide de deux outils différents. Le corpus Manuels FLE avec la plateforme Hyperbase web[2] qui permet de réaliser des bases hypertextuelles avec les textes qu'on lui fournit et combine des fonctions documentaires et statistiques pour décrire, caractériser, classer et interpréter les textes. La tokenisation, l'étiquetage morpho-syntaxique et la lemmatisation du corpus ont été effectués avec Spacy. Le corpus SCFLE a été constitué et analysé à l'aide de la suite de logiciels EXMARaLDA (Schmidt & Wörner, 2009) et CLAN (MacWhinney, 2000) ; la recherche de concordances sur EXMARaLDA-Exakt est facilitée par l'annotation de surface des MD et l'utilisation des expressions régulières (ReGex).

3.1. Les corpus

Les deux corpus numériques sont échantillonnés et constitués de transcriptions d'interactions orales en langue française de nature et de taille différentes.

Le corpus Manuels FLE[3], dont l'étendue est de 446.350 mots, contient les dialogues et les polylogues de divers ensembles pédagogiques de FLE[4]. Ces interactions, qui fournissent un modèle de la langue cible aux apprenants allophones, sont organisés en 6 classes correspondant aux 6 niveaux du CERCL. Le choix du corpus répond à plusieurs critères : les textes sont issus de manuels qui s'inscrivent dans l'approche actionnelle du CECRL, manuels de méthodes couvrant plusieurs niveaux, manuels de méthodes récentes, ouvrages conçus pour la préparation aux diplômes du DELF et DALF[5]. Ce dernier critère a été ajouté puisque la plupart des méthodes actuellement sur le marché ne couvrent pas tous les niveaux.

Le corpus SCFLE[6], qui comporte 15.861 mots, est constitué de dialogues entre

[2] Hyperbase a vu le jour en 1989 sous forme de logiciel grâce à l'œuvre d'E. Brunet au sein du laboratoire « Bases, Corpus, Langage » et n'a pas cessé d'évoluer. Depuis 2015, une version web a été développée par L. Vanni : http://hyperbase.unice.fr/.

[3] Pour cette étude nous n'avons exploité qu'une partie du corpus des manuels FLE accessible sur Hyperbase web qui contient aussi des monologues et dont la taille est de 614.025 mots. Ce corpus a été établi dans le cadre de nos recherches sur l'évaluation de(s) niveau(x) d'un texte en fonction des échelles du CECRL grâce au deeplearning (Ruggia et Vanni, 2022).

[4] Par manque de place, nous renvoyions à Ruggia (2021) pour les références des manuels utilisés.

[5] Diplôme d'Etudes de Langue Française, Diplôme Approfondi de Langue Française

[6] Le corpus SCFLE a été réalisé dans le cadre de notre thèse de doctorat (Rojas Madrazo, 2020) afin d'analyser le développement de l'apprentissage du FLE chez des apprenants en immersion à travers d'analyses longitudinales et contrastives. Il comporte au total 47.025 mots, dont 15.861 correspondant aux interactions des natifs ; 15.612 correspondant aux interactions entre apprenants; 8.767 correspondant à des monologues des natifs et 6.785 correspondant aux monologues des apprenants.

locuteurs natifs, étudiants de deuxième année de licence LEA (10 femmes et 4 hommes), dont l'échantillonnage a été effectué sur la base du volontariat. Les interactions portent sur des thèmes de communication proposés dans le CECRL (Conseil de l'Europe, 2001 : 45), tels que les loisirs, la nourriture et le transport. Les échanges ont été déclenchés après la lecture d'un texte bref en relation avec la thématique sélectionnée et de quelques questions.

3.2. Exploration du corpus et extraction des MD

Nous avons effectué une première exploration des deux corpus, consacrée au repérage des MD simples les plus fréquents et les moins fréquents dans chaque corpus. Cela nous a permis de distinguer 15 MD qui cumulent au moins 5 occurrences dans un de deux corpus. A partir de ces premiers résultats, nous nous sommes penchées sur les MD complexes comprenant au moins un des MD simples les plus fréquents, à savoir *ben*, *bon* et *du coup.* Nous avons également retenu ceux qui cumulent plus de 5 occurrences dans, au moins, un des deux corpus analysés, suivant le même critère que pour les MD simples. Il faut signaler que dans les cas où deux de ces trois marqueurs apparaissent ensemble ils sont regroupés sous le MD qui apparait en premier. Ainsi :

(1) *ben du coup* est classé comme MD complexe de *ben*.

(2) *du coup ben* est classé comme MD complexe de *du coup*.

3.3. Analyses

Étant donné que la taille des corpus est visiblement différente, nous avons normalisé les données en calculant le taux d'utilisation des MD par rapport au nombre total des mots des corpus, autrement dit en divisant le nombre total d'occurrences du MD observé par le nombre des mots du corpus où il apparait, multiplié ensuite par 100. Cette opération permet d'une part, d'obtenir la distribution statistique des MD dans les deux corpus et d'autre part, d'observer quelle est la proportion d'utilisation d'un MD par rapport aux autres MD analysés.

Le taux a donc été calculé pour les 15 MD simples retenus et pour les 15 MD complexes contenant *bon*, *ben* et *du coup*. Nous avons également déterminé le taux total de ces mêmes MD dans chaque corpus ; ce dernier montrera quelle est la proportion de la totalité de MD analysés.

Enfin, les résultats ont été soumis au test U-Mann Whitney pour deux échantillons non appariés, afin d'observer si la différence entre le taux d'utilisation des MD est significative. L'hypothèse nulle est qu'il n'y a pas de différences significatives entre les deux corpus.

4. Résultats

En ce qui concerne la distribution des fréquences, on constate des différences : dans le corpus Manuels FLE, les trois MD simples les plus utilisés parmi ceux qui ont été retenus sont *mais*, *donc* et *alors* (tableau 1), alors que dans le corpus SCFLE ce sont *mais*, *ben* et *donc* (tableau 1). Il est donc à noter que, le marqueur *mais* est le plus employé dans les deux corpus.

Les différences corroborées concernent aussi les MD simples les moins fréquents : *quoi*, *disons* et *genre* dans le corpus Manuels FLE et *tu vois, hein* et *disons* dans le corpus SCFLE. Là aussi on retrouve un même marqueur : *disons*, très rarement utilisé dans les deux corpus.

Tableau 1. Nombre d'occurrences et taux des MD dans les deux corpus.

MD simples	Corpus Manuels		Corpus SCFLE		MD complexes	Corpus Manuels		Corpus SCFLE	
	occurr.	taux	occurr.	taux		occurr.	taux	occurr.	taux
Nº de mots	446350		15861		Nª de mots	446350		15861	
MD s. analysés	6733	1,508	861	5,428	MD c. analysés	187	0,042	59	3,720
mais	2671	0,598	168	1,059	ah bon	89	0,020	1	0,006
donc	1473	0,330	122	0,769	mais bon	38	0,009	9	0,057
alors	1156	0,259	34	0,214	bon ben	13	0,003	0	0,000
là	801	0,179	11	0,069	bon mais	6	0,001	0	0,000
bon	687	0,154	10	0,063	bon alors	5	0,001	0	0,000
puis	584	0,131	47	0,296	ah ben	15	0,003	5	0,032
ben	470	0,105	125	0,788	ben alors	6	0,001	0	0,000
voilà	369	0,083	62	0,391	ben du coup	4	0,001	5	0,032
enfin	240	0,054	88	0,555	et ben voilà	2	0,000	1	0,006
hein	174	0,039	5	0,032	et puis ben	2	0,000	2	0,013
tu vois	65	0,015	5	0,032	donc du coup	5	0,0011	7	0,0441
du coup	59	0,013	99	0,624	et donc du coup	2	0,0004	0	0,0000
quoi	44	0,010	39	0,246	du coup ben	0	0,000	9	0,0567
disons	19	0,004	1	0,006	du coup ah	0	0,000	9	0,0567
genre	3	0,001	15	0,095	du coup euh	0	0,000	11	0,0694

Les résultats du taux d'utilisation de chaque marqueur (tableau 1) illustrent également les différences entre les deux corpus, puisque le taux est plus élevé dans le corpus SCFLE que dans le corpus Manuels FLE sauf pour les marqueurs *là, alors* et *bon.* Les différences les plus remarquables se trouvent entre l'utilisation des MD *ben*, *du coup* et *enfin* qui dépassent respectivement 0,5 points. Les différences les moins importantes concernent les MD *hein* et *tu vois*, très peu présents dans les deux corpus : *hein* est légèrement plus utilisé dans le corpus Manuels FLE (0,039 *vs* 0,032) alors que *tu vois* est plus fréquent dans le corpus SCFLE (0,032 *vs* 0,015). Ces disparités sont nettement visibles grâce au taux total : 1,5 MD tous les 100 mots (Manuels FLE) et 5,4 MD tous les 100 mots (SCFLE).

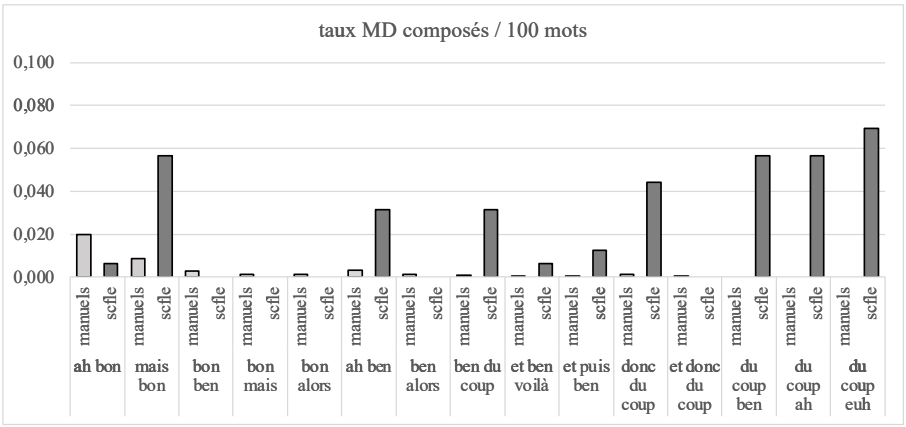


Figure 1. Taux MD complexes corpus Manuels FLE et corpus SCFLE.

L'écart entre les deux corpus est aussi marqué par les choix préférentiels des MD complexes (Figure1). Un éventail plus large ressort du corpus Manuels FLE, ce qui pourrait être dû à la taille du corpus, cependant le taux d'utilisation est encore beaucoup plus élevé (3,720 *vs* 0,042) dans le corpus SCFLE. En outre, le type de MD complexes les plus fréquents dans chaque corpus varie, ainsi dans les manuels les locuteurs privilégient les MD contenant *bon*, à savoir *ah bon, mais bon* et *bon ben*, tandis que ce sont les MD composés par *du coup* : *du coup euh, du coup ah* et *du coup ben* les préférés par les locuteurs natifs. Enfin, les MD complexes *et ben voilà*, *et puis ben*, *ben bon* sont les moins utilisés dans les deux corpus. Malgré les différences trouvées, les résultats du test U Mann Whitney montrent qu'elles ne sont pas significatives statistiquement ni dans le cas des MD simples ($z = -.772$; $p > 0.05$) ni dans le cas des MD complexes ($z = -.628$; $p > 0.05$).

5. Conclusion

L'étude exploratoire des corpus Méthodes FLE et SCFLE a démontré que la fréquence d'utilisation, ainsi que les choix préférentiels des MD simples et complexes sont plus élevés lors des interactions orales des locuteurs natifs français que lors de celles proposées comme modèles discursifs dans les manuels aux apprenants de FLE.

Ces premiers résultats, qui se doivent d'être corroborés par l'étude d'autres grands corpus, nous amènent néanmoins à nous interroger sur la typologie et les emplois des marqueurs discursifs dans les interactions orales présentes dans les manuels FLE. Ainsi, cette étude n'est à considérer que comme une première étape d'une recherche qui vise à analyser les marqueurs discursifs aussi bien d'un point de vue quantitatif que qualitatif, afin de voir si les modèles discursifs

destinés aux apprenants de FLE reflètent les caractéristiques du français oral, notamment du français spontané.

Bibliographie

Bazzanella C. (2006). Discourse markers in Italian: Towards a «compositional» meaning. In Fischer, K. (ed.) *Approaches to discourse markers*. Elsevier.

Chanet C. (2004). Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. *Recherches sur le français parlé*, 18: 83-107.

Conseil de l'Europe. (2001). *Cadre Européen Commun de Référence pour les Langues*. Unité de politiques linguistiques. Disponible sur : https://rm.coe.int/cadre-europeen-commun-de-reference-pour-les-langues-apprendre-enseigne/1680a4e270

Dostie G. (2004). *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*. De Boeck, Duculot.

Dostie G., et Pusch C. D. (2007). Présentation. Les marqueurs discursifs. Sens et variation. *Langue française*, 2(154): 3-12.

MacWhinney B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs.* Lawrence Erlbaum Associates Publishers.

Mayaffre D. (2010). *Vers une herméneutique matérielle numérique. Corpus textuels, logométrie et langage politique*. Habilitation à Diriger des Recherches. Université Nice Sophia Antipolis.

Rodriguez Somolinos A. (2011). Présentation. Les marqueurs du discours – Approches contrastives. L*angages,* 4(184): 3-13. Disponible sur : http://cairn.info/revue-langages-2011-4page-3.htm

Rojas Madrazo M. (2020) *Stratégies de communication, fluidité et lexique en production orale. Étude longitudinale d'apprenants en immersion*. Thèse de doctorat. Université Savoie Mont Blanc.

Ruggia S. (2021). La lecture contrôlée et assistée par l'analyse statistique des données textuelles : comment et pourquoi interroger un corpus numérique ?, *Le français dans le monde, Recherches et Applications, Langues et pratiques numériques : nouveaux repères et nouvelles littératies en didactique des langues ?,* 69: 84-100.

Ruggia S. et Vanni L. (2022). DeepFLE : la plateforme pour évaluer le niveau d'un texte selon le CECRL, *Dialogues et Cultures*, 66 : 235-254.

Schmidt T., et Wörner K. (2009). EXMARaLDA–Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4): 565-582.

Waltereit R. (2007). A propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de bon ben et enfin bref. *Langue française*, 2(154): 94-109. Disponible sur : http://www.cairn.info/revue-langue-française-2007-2-page-94.htm

Assessing Corporate Social Responsibility Communication on Twitter: A Composite Indicator Approach

Camilla Salvatore[1], Silvia Biffignandi[2], Annamaria Bianchi[3]

[1]University of Milano-Bicocca – c.salvatore4@campus.unimib.it
[2]CESS – biffisil@teletu.it
[3]University of Bergamo – annamaria.bianchi@unibg.it

Abstract
The aim of this paper is to provide a framework for the construction of a composite indicator (CI) to characterize Corporate Social Responsibility (CSR) online communication. To this purpose we focus on the firms included in the Dow Jones Industrial Average index and their Twitter accounts. The proposed CI results from the combination of two macro-dimensions: communication *content* and communication *modality* (information and media richness). In order to identify the content, we use the Structural Topic Model (STM) which allows to discover and link the topics to CSR dimensions. We discuss the construction of simple indexes, and we address aggregation strategies. The proposed composite indicator may help researchers and policy makers to monitor the behaviour of companies with reference to sustainable development and the Agenda 2030. It also contributes to the advancement of methodological knowledge for the construction of socio-economic composite indicators using textual data.

Keywords: Socio-Economic Indicators; Corporate Sustainability; Sustainable Development; Social Media; Structural Topic Modeling.

1. Introduction

In the path towards a more sustainable world, businesses play a crucial role in addressing socio-economic matters contributing to the development of local and global communities (Imaz and Eizagirre, 2020). Socially responsible behaviours of businesses are linked to the concept of Corporate Social Responsibility (CSR). This refers to the implementation of activities aiming at the improvement of firms' reputation and at positively impacting the society (Caroll, 1991).
The availability of social media (SM) data provides an excellent opportunity to investigate the online communication of CSR activities. Indeed, listening to the online communication is useful to researchers and policy makers in order to monitor the behaviour of the business with reference to the implementation of sustainable development and with respect to the Agenda 2030. Despite its

importance, the study of online business communication with respect to sustainability is still a relatively under-examined and emerging topic (Chae & Park, 2018; Araujo & Kollat, 2018). Further, given its multi-faceted nature, measuring CSR activities is naturally related to the use of composite indicators, which allow to summarise complex or multi-dimensional phenomena (Dahlsrud, 2008). In this respect, it is worth noticing that the construction of composite indicators (CI) based on textual data is a new spreading research area.

While SM-based CI have been developed to study various social aspects according to citizens' vision, such as well-being (Iacus et al., 2020), social tensions[1], and consumers' confidence on the economy (Daas and Puts 2014; Istat's Social Mood on Economy Index[2]), as far as we know, few studies have investigated the online behaviour of firms with respect to sustainable development. This paper aims to fill this research gap, with a twofold purpose, entailing both a methodological (construction of indexes and indicators) and a substantial contribution (analysis of CSR communication). We develop a framework to construct a composite indicator (CI) resulting from the analysis of tweets. Such indicator is built in two steps and derives from the combination of two macro-dimensions. The first one refers to the communication *content* in tweets which is identified applying the Structural Topic Model which allows to discover and link the topics to CSR dimensions. The second one refers to communication *modality* (media richness).

To develop the CI, we first select the firms included in the Dow Jones Industrial Average index and we identify the topics discussed and their relationship to CSR dimensions. We do not classify the text into pre-defined categories, but we study the underlying properties of the text in order to get signals from the data and to identify the topics the text is about. To this purpose, we apply the Structural Topic Model (STM) algorithm, which allows to estimate the model including document-level metadata (time and sector). Topic proportions contribute to the content dimension, while tweets characteristics contribute to the modality dimension. The proposed CI is built in two steps. First, we combine simple indexes to generate a CI for each macro-dimension following a compensatory approach and then, we combine the two macro-dimensions through the Mazziotta-Pareto non-compensatory approach (De Muro et al., 2011). We generate a CI aggregated at sectoral level. The resulting CI measure the online communication effort of businesses about CSR. From the analysis of the CI we expect to find that the communication of firms in the services sector will be more effective than the communication in other sectors.

The reminder of the article is the following. The methodology and the data are

[1] Social Tension Indicator: link

[2] https://www.istat.it/en/experimental-statistics/experiments-on-big-data

presented in Section 2 and 3, respectively. Results are discussed in Section 4 and conclusions are drawn in Section 5.

2. Methods

2.1. The Structural Topic Model (STM)

In order to identify the content of unstructured textual data, a common approach is to implement topic modelling (TM). It is an unsupervised learning technique which allows to study the underlying properties of a text in order to discover the topics discussed and get signals from the data (Vayansky & Kumar, 2020). Among the different algorithms to implement TM, we select the STM which was originally designed to analyze open-ended survey questions, and which is becoming increasingly popular due to the possibility of estimating models including document-level metadata and, thus, characterizing the relationship between topics and metadata (Roberts et al., 2016).

In the following, we briefly introduce the STM algorithm. For more details, please refer to Roberts et al. (2016). Figure 1 represents the model in plate notation. A topic is defined as a mixture over words and a document as a mixture over topics. In STM, document-metadata influences two components of the model, the topical prevalence that is defined as the proportion of the document that is associated to a topic, and the topical content that refers to the usage rate of word in a topic. Here we focus only on topical prevalence covariates. In particular, we estimate the effect of time and sector on the discussion proportion of topics as part of a larger application-oriented study.

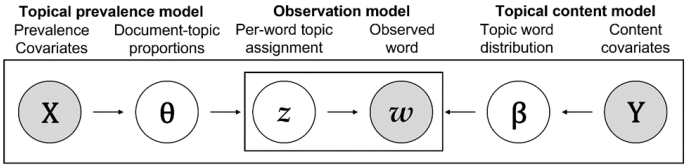


Figure 1. Structural Topic Model. Source: Amended from Roberts et al. (2016).

For our analyses, we use R and, in particular, the `stm` package (Roberts et al., 2019) to estimate the model and the `quanteda` package (Benoit et al., 2018) to clean and prepare the data.

2.2. The Composite Indicator Approach

We propose to construct a CSR-CI which summarizes two aspects of the online communication, the *content* and the *modality*. Aggregation follows a two-steps procedure. First, we derive two indexes for content and modality, respectively, using a compensatory approach. Next, we aggregate them using the Mazziotta-Pareto index (MPI), a non-compensatory approach which rewards the unit that,

for the same average level, has a greater balance in elementary indicators.
Content refers to the aspects firms communicate online. Through STM we identify several topics, and we link them to the following four CSR macro-dimension: social, economic, environmental, mixed-general CSR, and non-CSR. For each tweet, we consider the overall proportion of text devoted to CSR, and then we aggregate these proportions at sectoral level taking the arithmetic mean.
Modality refers to the way firms communicate. This is an important aspect for the communication to be effective and to engage with customers and stakeholders. In this respect, each tweet can contain hashtags (defining the topic of posts and allowing users to associate the tweet with all other tweets using the same identifying hashtags), mentions (engaging with other users), media (photos), and links (to external webpages). We expect that the higher the media richness, the more effective the communication will be (Araujo & Kollat, 2018). For each tweet, we create an overall indicator of modality of communication by computing the sum of individual indicators about the presence of hashtags, mentions, media, and links, and then we aggregate this index at sectoral level by taking the arithmetic mean.
Finally, we aggregate the content and modality indexes to build a CSR communication index (CRS-CI), measuring both CSR content and effectiveness of communication. To compute the MPI, we standardize the data as in (1), where *i* indexes the sector and *j* the indicator. *M* and *S* refer to mean and standard deviation of the content and modality indexes.

$$z_{ij} = 100 + \frac{(x_{ij} - M_{xj})}{S_{xj}} \cdot 10. \quad (1)$$

Then, we compute the MPI as in (2) given the positive polarity[3] of the indicator, where *z* refers to the standardized data as in (1) and M_{zi}, S_{zi}, cv_{zi} denote the mean, standard deviation, and coefficient of variation of the normalized values for unit *i*, respectively.

$$MPI_i = M_{zi} - S_{zi} \cdot cv_{zi}. \quad (2)$$

Figure 2 summarizes the aggregation approach described above.

[2] The polarity of an indicator refers to the sign of the relationship between the indicator and the phenomenon to be measured

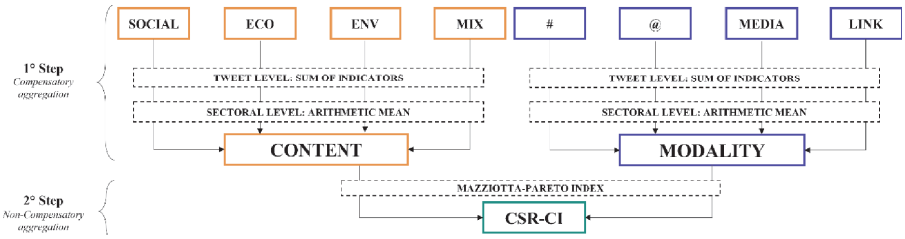


Figure 2. Composite Indicator aggregation strategy.

3. Data

To carry out our analysis, we selected the firms included in the Dow Jones Industrial Average index, i.e., a stock market index that measures the performance of the 30 largest US listed companies as of the last composition change in August 2020. We retrieved the full list of firms, jointly with the corresponding activity sector from Bloomberg. With respect to sectors classification, Bloomberg adopts the Global Industry Classification Standard (GICS) developed by MSCI and S&P Dow Jones. Next, we identified the official Twitter accounts of the companies. Given that companies may have several Twitter accounts, we focus primarily on CSR accounts and, in case these are not available, on the news or multipurpose ones. The objective is to reduce the noise (no-CSR tweets) in the data. Apple and Walgreens Boots Alliance turned out not to have a Twitter account, thus, leading to the inclusion of 28 firms and 42 different accounts (18 CSR, 5 news-type and 19 multipurpose) in the analysis. Looking at the sectoral composition, most of the companies operates in the Technology sector (21.4%), followed by the Financial, Health Care and Industrials sectors (14.3% each), Consumer Discretionary and Staples (10.7% each), Communication and Energy & Materials (7.14% each). The total number of messages retrieved is 25,148 with a mean of 2,096 messages per month.

4. Results and Discussion

We present a summary of the results about topic modelling. More details are available on a forthcoming paper where covariates effects are discussed. Hereafter, we focus our attention on the development of the composite indicator. We identified 47 topics, 36 of which related to CSR activities. The majority of the topics (43%) relates to the social dimension (inclusiveness, equality, well-being etc.). This is followed by the mixed-general CSR communication (17% - mainly about the sponsorship of sports and cultural events), the economic and the environmental one (13% and 4% respectively).

From the sectoral analysis, it is possible to identify some clusters, showing similar communication content. The social dimension is prevalent in all sectors, but especially in the Health Care, Technology, and the Communication ones. The

economic dimension is specific to the Financial sector, together with the Industrial one. Environmental communication is more important for the Energy & Materials sector. The mixed-general CSR communication is prevalent in the Industrials and Communication sectors.
Based on the topic model output, we construct and present the composite CSR communication indicator in Table 1.

Table 1. MPI for CSR communication.

Sector	CSR-CI
Health Care	109.96
Financials	105.08
Technology	104.73
Communications	98.03
Energy & Materials	95.49
Industrials	94.69
Consumer Staples	94.02
Consumer Discretionary	88.88

The best and worst performing sectors in terms of CSR communication are Health Care (HC) and Consumer Discretionary respectively. The CSR-CI is above 100 for HC, that shows an effort toward social responsibility based on online communication. Also, Financials and Technology sectors give an indicator higher than 100 suggesting a rich and effective online communication. The performance of other sectors is quite similar, except for Consumer Discretionary where the CSR-CI is equal to about 89. The CSR-CI index by sector confirms the hypothesis that services sectors, especially HC and financial services, are more committed toward sustainability online communication. Indeed, the implementation of CSR highly needs to go through the website communication, whereas the Consumer Discretionary sector does not need an intensive online communication probably because other ways to communicate sustainability are available when offering a consumer product (such as information on the package).

5. Conclusions

The availability of new data sources, like social media, represent a big opportunity to study new socio-economic phenomena. In this paper, we address the methodological challenge of building a composite indicator from textual data and we also study the online communication of CSR to get new insight about business behavior and commitment in sustainable development.
More details about topic modeling results and their relationship with Sustainable Development Goals (SDGs) are available in a forthcoming paper.

However, the analysis of Twitter data is not error free and practitioners should be aware about quality aspects and different source of errors (Salvatore et al., 2020). Limitations of the present study include the sensitivity of topic model to analysts' judgements at the various steps.
Future developments include the construction and validation of dictionary about CSR dimensions which can be useful to replicate the analysis on new data in a faster and user-friendly way. Indeed, the *content* dimension can be computed using different natural language processing methodologies. In addition, also the integration of social media-based CI with survey-based ones is of interest, and we plan to investigate this aspect in a future study.

References

Araujo T. and Kollat J. (2018). Communicating effectively about CSR on Twitter: The power of engaging strategies and storytelling elements. *Internet Research*, 28(2), 419-431.

Benoit K., Watanabe K., Wang H., Nulty P., Obeng A. and Müller S. M. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. doi: doi: 10.21105/joss.00774

Carroll A. (1991). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons*, 34(4), 39-48.

Chae B. and Park E. (2018, 6 28). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability*, 10(7), 2231.

Daas P. J. H. and Puts M. J. H. (2014). Social media sentiment and consumer confidence, *European Central Bank Statistics Paper Series*, No. 5.

Dahlsrud A. (2008, 1 1). How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management*, 15(1), 1-13.

De Muro P., Mazziotta M. and Pareto A. (2011). Composite indices of development and poverty: An application to MDGs. *Social indicators research,* 104(1), 1-18.

Iacus S. M., Porro G., Salini S. and Siletti E. (2020). An Italian composite subjective well-being index: The voice of Twitter users from 2012 to 2017. *Social Indicators Research*, 1-19.

Imaz O. and Eizagirre A. (2020). Responsible innovation for sustainable development goals in business: An agenda for cooperative firms. *Sustainability*, *12*(17), 6948.

Roberts M. E., Stewart B. M. and Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111, 988-1003.

Roberts M., Stewart B. and Tingley D. (2019). Stm: An R package for structural topic models. Journal of Statistical Software, 91(1), 1-40.

Salvatore C., Biffignandi S. and Bianchi A. (2020). Social Media and Twitter Data Quality for New Social Indicators. *Social Indicators Research*. doi:10.1007/s11205-020-02296-w

Vayansky I. and Kumar S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582.

The use of IRaMuTeQ for longitudinal analysis. A study on postgraduate course participants' reflections

Sara Santilli, Stefano Sbalchiero

University of Padova – sara.santilli@unipd.it
stefano.sbalchiero@unipd.it

Abstract
The analysis of textual data is a complex process and requires clear thinking on the part of the analyst. However, several shortcomings can hinder the research analyst during the process, leading to inconsistencies, especially when qualitative and manual content analysis is applied. This article reflects the use of a quanti-qualitative approach to content analysis by Iramuteq software and its usefulness in identifying categories of analysis, automatically extracting topics and analyzing their differences at a longitudinal level. Specifically, a study was conducted considering the reflections of 43 participants in a postgraduate course aimed at journalism professionals promoting inclusive language and increasing attention to information manipulation phenomena such as fake news. Reflections from each weekend of the course (nine in total) were collected for each participant. The software Iramuteq was used to code, analyze and extract the main topics considering the longitudinal level. Results show that as a quanti-qualitative analysis of textual data, the approach is a valuable means to promote the robustness of quanti-qualitative research also at the longitudinal level and to highlight changes in participants' reflections during activities such as courses and interventions, or counselling and therapy sessions.

Keywords: reflection analysis, quali-quantitative analysis, longitudinal textual data, IRAMUTEQ.

1. Introduction

There is a growing aspiration that higher education transforms students into 'critical reflective thinkers' able to cope with a rapidly changing world (Harvey & Knight, 1996). A core aspect in different definitions of critical reflective thinking is creating meaning or conceptualization from a learning experience that enables a learner to see things differently (Seal and Cann, 2000). A learner may try to create meaning from the learning experience by reframing the experience. According to Mackinnon (1993) and Smith (1994), reframing involves re-examining the problem (learning experience) from one or more

theoretical platforms. Schon (2010) defined reflecting on the action as the thinking after a training event. Thinking back on what we have done to discover how the knowledge we put into action may have contributed to an unexpected outcome. In addition to producing definitions of reflection, writers have discussed how reflection can be facilitated. For example, Laurillard (1993) argues that reflection can therefore be facilitated if teachers discuss and negotiate with students what the goals might be, give feedback (adopt a student's perception and enable them to experience it from different perspectives) and give students opportunities to reflect on their experiences and to internalize them. This process of discussion and negotiation between teacher and student may be defined as a reflective dialogue (Brockbank & McGill, 1998). In the present study, we wanted to analyze through quali-quantitative procedures the reflective process of a group of 43 participants of a Higher Education course "The passion for truth, how to inform by promoting an inclusive society". Participants, communication and journalism professionals, were involved in a reflective process. Professor asked them from time to time not only to study the materials provided for each lesson and to verify the understanding of the same but also to answer two questions: 'the lesson made me think about....' and 'What struck me most about this lesson is....', to converge thought on the contents, stimulate links with personal experience and/or others, seek explanations and analysis, search for meaning. The fact of asking them lesson by lesson wanted to stimulate a recursive and cyclical thought, aimed at expansion, openness and deepening, at the creation of a broader conceptual construction of understanding reality with the idea of promoting a critical conscience and reflective processes, a sense of social responsibility for what is happening, aspirations for a better world, stimulating active participation and a process of co-construction (Nota and Rossier, 2015). So with the reflective activities proposed by the Higher Education course, we wanted to involve the participants in moments aimed at understanding the present, the neoliberal culture and the myths promoted by it, the simplistic and reductive readings of human development and realization together with hedonistic, narcissistic and cynical visions of reality. The present study aims to investigate the themes underlying the reflections produced by the courses during the meetings and analyze the differences at the longitudinal level.

2. Corpus and method

The corpus is composed of the answer provided by the course participants organized by the longitudinal variable, i.e. from meeting Nr.1 to meeting Nr.9. The corpus was pre-processed using IRaMuTeQ software (Ratinaud 2014; Sbalchiero, 2018). The normalization consisted in replacing uppercase with lowercase letters, like punctuation, numbers and stop words have been removed because they are not significant in terms of content. Moreover, the pre-processing

steps helped reduce redundancy and provide homogeneity among forms. In this regard, the lexicometric measures (Tab 1.) clearly indicate that a statistical analysis of textual data is plausible for this corpus (Lebart et al. 1998), taking into consideration the ratio between hapax and type is less than 50% (42,86%).

Table 1. Lexicometric characteristics of the analyzed corpus

Number of texts	9
(N) word-tokens	204007
(V) word-type	11443
(V1) hapax	4904
(V/N)*100 = Type/Token Ratio	2,4
(V1/V)*100 = Percentage of hapax	42,86

By applying Reinert's method (1983) to our corpus, made available in the R-based version of IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires) (Ratinaud, 2014a), we extracted a series of topics, or what Reinert (1993) calls "lexical worlds". If the procedure has the advantage of identifying the lexical worlds contained in a corpus by reducing the biases of merely qualitative encoding, on the other side it permits to detect topics that are more likely to appear as clues disseminated among the answers of participants. The procedure involves several successive stages: after importing and pre-processing the corpus, the next step consists of the automated identification of minimal units called elementary context units (ECU) following two empirical criteria: the length of the ECU (the maximum number of words they can contain, 40 words by default), and punctuation marks (if an ECU ends at 38 with a dot, punctuation is respected). Applied to the corpus analyzed, this led to the identification of 5,647 ECU containing a mean of 36 words each. Then, the algorithm identifies co-occurrences in each ECU by constructing a contingency matrix of *words* x *units* organized as a binary table (absence vs presence of a given word in the portion of texts). This matrix provides the basis for analyzing similarities between ECUs, which are summarized using descending hierarchical cluster analysis based on a sequence of correspondence analysis that split binary matrix in two clusters at each step: the "clustering procedure hierarchically identifies the factors (clusters) that best represent a lexical world from the distance of the χ^2 between the classes […]. The classification outcome is a set of classes, each representing a lexical world, i.e. a set of ECUs that include words relevant to the class. We deduce that ECUs are all the more similar, the more words they have in common. The list of the most meaningful words that best represent a lexical world is identified by

associating the χ^2 between words and classes" (Sbalchiero, 2018, p. 203-204). For the main aim of the present paper, the results produced by the algorithm can be used to assess the classes' association with the number of meeting from Nr. 1 to Nr. 9. Therefore, to verify the main topics that emerged and their distribution over time, we examined the longitudinal dimension of the lexical worlds using graphical representations (Ratinaud, 2014b) referred to the proportion of classes by meeting and the intensity of classes by the meeting.

3. Results

The procedure identified seven classes (or lexical worlds), as the dendrogram shows (Fig. 1), and it works pretty well because almost 9 out of 10 ECUs have been classified (87%). Consequently, the words characterizing a class are semantically associated and not only do they refer to the same issue, but they also make the interpretation of each class.

Specifically, the first class refers to inclusion and social sustainability; it is characterized by words that bring to the center inclusion and social sustainability in their fullest and most current concepts, of combating inequalities and discrimination, of enhancing heterogeneity; class 2 concerns the words about the importance of journalistic activities to bring to everyone and to all information and knowledge that describe reality; class 3, journalism as a process of approaching the truth: collects words and phrases related to the idea of journalist and communication expert as a professional who seeks the truth, albeit with difficulty, considering the complexity, the context, the environmental variables in addition to the individual ones; he makes deeper analyzes, avoids reductivism and simplism, assuming a more significant social role for democracy; class 4, the different places where journalism operates, from the printed press, in social media, in the areas from where the stories are born; class 5, accuracy and inclusive language: it is characterized by words that recall the importance of accuracy and a careful and inclusive language, which avoids labels and stereotypes, in order to avoid the spread of fake news and to counter forms of hatred; class 6, decent work: contains the words concerning the idea of decent work as it is based on fair compensation, safety, health protection and trade union protection; class 7, concern the reflections of the participants on the specific topic of each meeting. The results highlighted that the topics detected in our corpus contain the main themes deepened during the course.

Finally, the classes were presented using some graphical representations to compare and contrast the topics over the different meetings. On the one hand, insofar as a given meeting can be represented as the sum of all the ECUs associated with that meeting, we can reproduce each meeting as a distribution of the proportion of classes that characterize it (Fig. 2).

The width of the bars is proportional to the number of ECUs in a given meeting,

and their height represents the ECUs in that meeting, so we can see the distribution of the classes from a longitudinal perspective. This representation, in fact, reveals the trend of the topics over the meeting. While classes 7 and 2 (which refer to reflections of the participants on the specific topic and the importance of journalistic activities to bring to everyone and to all information and knowledge that describe reality) were typical of the first meeting, a rising trend can be seen especially for the class 3, related to the concept of journalism as a process of approaching the truth. The others can be defined as "evergreen topics" because they are present across the different meetings, even if to a different extent. If we wanted, instead, to focus on the general topic at each meeting, we could use the contributions of χ^2 expressed by the modalities of the 'meeting' variable. If a topic is discussed more during a given meeting, then the positive differences and the threshold for the significance of χ^2 will indicate an association between meeting and semantic class (Fig. 3).

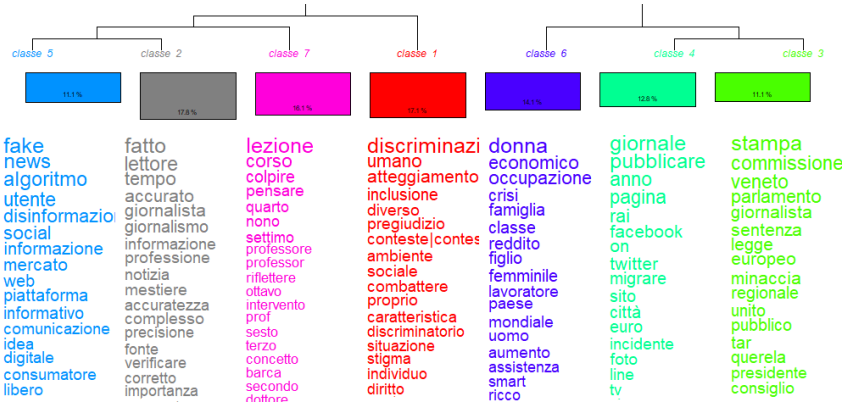


Figure 1. Cluster dendrogram and most relevant words for each lexical word (by descending order of χ^2 between words and classes)

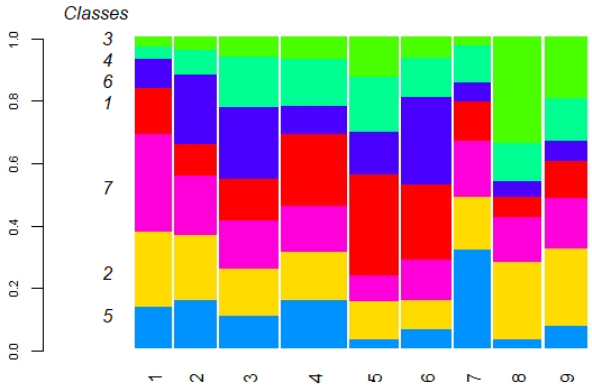


Figure 2. Proportion of classes by year based on the sum of the ECUs aggregated by class in each meeting (from meeting Nr. 1 to Nr. 9)

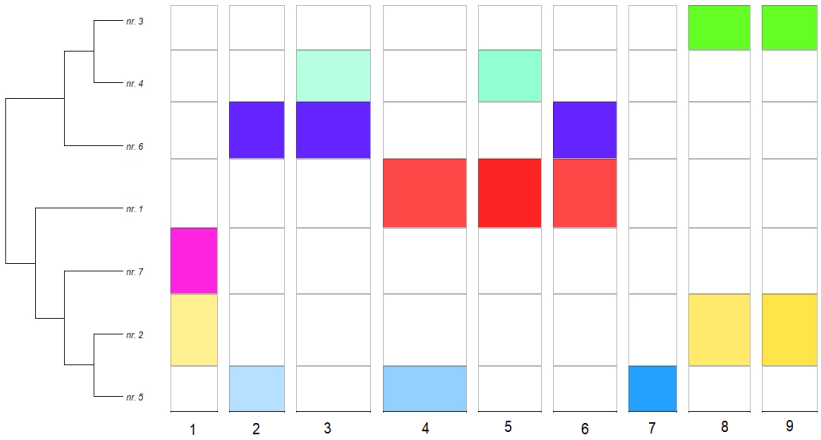


Figure 3. Intensity of classes by meeting (from meeting Nr. 1 to Nr. 9)

In this case, the height of lines of each class is proportional to the class dimension in terms of the number of ECUs it contains (class 3 and 5 is the smallest), and the width of the cells is proportional to the frequency of ECUs in a given year. The intensity of colour, i.e. proportion association between class and meeting, is based on the threshold for the significance of χ^2: a p-value ≥ 0.05 for the white boxes, and <0.0001 for the darkest boxes, which are the most significant. The shades of colour vary between these two extremes, so class Nr. 5 increased its intensity because it was present, with different intensity, in meeting Nr. 2, Nr. 4 and Nr. 7. Similarly, topic Nr. 2 is associated with central meetings, and its intensity is more significant in the fifth meeting.

4. Conclusions and discussion

The present study aimed to analyze topics and show their changes in a longitudinal perspective using a variable capable of highlighting the order of a series of meetings. On one side, the classes that emerge from the interviews of the journalists who participated in the course bring out an idea of inclusion in harmony with recent literature, identify barriers and hindering factors that the context manages to pose to a quality life for all and to human promotion. On the other hand, the course specifically wanted to stimulate a conceptual mastery that considered the complexity and interdependence of events, where a local event can have repercussions on a global level and vice versa, and where there is a close intertwining between our lives and the lives of those who live in other continents, between the destiny of humanity and that of the planet (Deleuze-Guattari, 2017). The course wanted to train people to take a holistic vision, which has its roots in responsibility and commitment in a professional way (Arendt, 2003), imagining new futures and societies, which can only be inclusive,

sustainable, based on social justice, to which everyone can contribute, to which all disciplines and professions are called to make their contribution.
On the other side, from a methodological point of view, the analysis helped interpret the peculiarities that specific topics assumed during the meetings. In this paper, we tried to analyze how the reflective process evolves in a group of communication and journalism professionals, participants in an educational activity, able to consider more frequently the issues of complexity, accuracy, critical thinking, decent work, inclusion and sustainability, as tools of struggle, with courage, to inequalities, discrimination, hatred.

References

Brockbank A. and McGill I. (2007). *Facilitating reflective learning in higher education*. McGraw-Hill Education (UK).
Harvey L. and Knight P. T. (1996). *Transforming Higher Education*. Open University Press, Taylor & Francis, 1900 Frost Road, Suite 101, Bristol, PA 19007-1598.
Laurillard D. (1993). *Rethinking University Teaching: Rethinking University Teaching: a Framework for the Effective Use of Educational Technology*. Routledge.
Lebart L., Salem A. and Berry L. (1998). Exploring textual data. Boston: Kluwer Academic Publication.
Mackinnon A. M. (1987). Detecting reflection-in-action among preservice elementary science teachers. *Teaching and teacher education*, *3*(2), 135-145.
Ratinaud P. (2014a). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [software, Version 0.7 alpha 2]. Retrieved from http://www.iramuteq.org
Ratinaud P. (2014b). Visualisation chronologique des analyses ALCESTE : application à Twitter avec l'exemple du hashtag #mariagepourtous, In Actes des 12eme Journées internationales d'Analyse statistique des Données Textuelles, JADT 2014, Paris, pp. 553-565.
Reinert M. (1983). Une methode de classification descendante hierarchique: application a l'analyse lexicale par contexte. Les Cahiers de l'Analyse des Données, 8(2), 187–198.
Reinert M. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars, Language et Société, 66, 5–39.
Sbalchiero S. (2018). Finding topics: a Statistical Model and a Quali-Quantitative Method. In A. Tuzzi, Tracing the Life-Course of Ideas in the Humanities and Social Sciences. Springer, 189-210.
Schon D. A. (2010). Educating the reflective practitioner: toward a new design for teaching and learning in the professions. *Australian Journal of Adult Learning*, *50*(2), 448-451.
Seale J. K., & Cann, A. J. (2000). Reflection online or offline: The role of learning technologies in encouraging students to reflect. *Computers & Education*, *34*(3-4), 309-320.
Smith M. (1994). *Local education: Community, conversation, praxis*. McGraw-Hill Education (UK).

Dimensionality Reduction of Unstructured and Network Data for Stance Detection

Andrea Sciandra

University of Modena and Reggio Emilia – andrea.sciandra@unimore.it

Abstract
The idea behind this work stems from the participation in some shared tasks concerning stance detection in NLP conferences. In these competitions, participants tried to develop the best stance prediction system for 'favor', 'against', and 'none' categories on selected topics, according to messages and relationships among users of a social networking site. Thus, the data available consisted of textual and network data. The teams we collaborated with used dimensionality reduction methods for network data, through a Multidimensional Scaling. On the other hand, the approach towards textual data involved different methods of feature extraction, without paying particular attention to dimensionality reduction for unstructured data. In this paper we show the empirical results of a two-step strategy to obtain lower-dimensional textual data relying on text mining techniques and principal component analysis. The results show levels of accuracy comparable to classical feature extraction techniques and to the best task models, despite using a much smaller number of predictors.

Keywords: Dimensionality Reduction, Stance Detection, Social Media, PCA, MDS, Multinomial Regression.

1. Introduction

Stance detection is the task of automatically classifying whether the author of a text is in favor, against, or neutral towards a given target. The target could be a person, an organization, a government policy, a movement, a product, etc. (Mohammad et al., 2017). Stance detection could be useful for several undertakings, such as market analysis, opinion surveys, predictions for elections, policy-making, media monitoring, and security (Küçük and Can, 2020). In fact, a stance detection could support the automatic identification of people's extremist tendencies on the one hand, but it could also be employed by authoritarian governments as a tool to control their citizens (Lai et al., 2021).

The idea behind this work stems from our participation in some shared tasks concerning stance detection in NLP conferences (Evalita 2020[1] and IberLEF 2021[2]). In these competitions, participants tried to develop the best stance

[1] www.evalita.it/evalita-2020
[2] sites.google.com/view/iberlef2021

prediction system for 'favor', 'against', and 'none' categories on selected topics, according to messages and relationships among users of a social networking site (Twitter). The shared tasks dealt with the stance detection of Italian tweets about the Sardines movement (SardiStance) and Spanish or Basque tweets about the Antivaxxers movement (VaxxStance). In particular, SardiStance was the first shared task regarding Stance Detection in Italian tweets, where the organizers collected posts about the Sardines movement and invite at automatically detecting their stance (Cignarella et al., 2020).
The data available consisted primarily of textual and network data. In both cases, the teams we collaborated with used dimensionality reduction methods for network data, through a Multidimensional Scaling (MDS) applied to distance matrices among social media users. On the other hand, the approach towards textual data involved different methods of feature extraction, without paying particular attention to dimensionality reduction for unstructured data. In this paper we show, regarding the SardiStance tasks, the empirical results of a two-step strategy to obtain lower-dimensional textual data relying on text mining techniques and principal component analysis (PCA) of the reduced document term matrix (DTM).

2. Data processing

The data available for these competitions included the text of the posts, information about the posts (number of retweets, number of favors, etc.), information about the authors (number of followers, number of friends, etc.), and information about their social networks (friends, replies, retweets, and quotes' relationships). In this contribution we focused particularly on post texts and social networks of friendship and retweeting, therefore dealing essentially with unstructured data and network data.
Our strategy for limiting the dimensionality of unstructured data is based on two steps. The first step involves treating the textual corpus through text mining techniques for features reduction. Specifically, the processes of text normalization, lemmatization, less frequent terms removal, and normalized TF-IDF calculation were applied. The text normalization included recognition of URLs and emoticons, stopwords removal, and slang/jargon correction through the `TextWiller` R package (Solari et al., 2019). Following normalization and lemmatization, lemmas with fewer than 30 occurrences were removed. In this way, we reduced the initial corpus containing 14815 types and 89183 tokens to a reduced corpus containing 254 types and 23338 tokens.
The second step involves a dimensionality reduction through statistical techniques. When we have data with a large number of dimensions, one goal may be to project this data into a lower dimensional subspace, while trying not to lose valuable information about the original variables. One way to achieve

this reduction is through the selection of variables, also called feature selection. Another way is through the creation of a reduced set of linear transformations of the initial variables. The creation of these composite features, using projection techniques, is often referred to as feature extraction. Principal component analysis (PCA) is a technique designed to derive, from a set of correlated numerical variables, a smaller set of orthogonal variables. The reduced set of linear orthogonal projections is obtained by linearly combining the original variables. Based on these considerations, we extracted the principal components from the reduced DTM, by first standardizing the normalized TF-IDFs. Following the PCA, the first 100 resulting dimensions were selected (cumulative proportion of variance: 50.7%) and used as features to predict the stances.

Conversely, with respect to network data, our approach to reduce the dimensionality of the data was based on the MDS applied to distance matrices among social media users (minimum path of graphs of 'friendship' ties and ties generated by retweets). In this work we consider only friendship and retweeting relationships because reply and quote graphs showed few relationships. So, for each network, a distance matrix among Italian twitterers was computed. We defined the distance among two users as the shortest path, forcing the graph to be undirected. The distance matrix was then projected into a Euclidean space through a MDS (Klimenta and Brandes, 2012). We expected the users to be partially polarized in clusters within the networks and, consequently, the largest dimensions of the MDS should polarize the stances. An exploratory data analysis confirmed our expectation: for instance, in Figure 1 (top) we show the scatter plot of dimensions 1-2 and 1-3 for the Friends Network. The first dimension clearly discriminates the three stances, in particular 'favor' and 'against'. Therefore, we decided to retain the four main dimensions for each of the two networks. Finally, we added the 8 dimensions extracted from the MDS coordinates for friendship and retweeting networks to PCA features.

Inspecting the bottom part of Figure 1, we must note that the first dimensions derived from PCA do not seem to discriminate efficiently among stances. Anyway, we expected a similar result due to the unstructured data, but we also expected that the exploitation of several components would be helpful to classify stances with substantial accuracy.

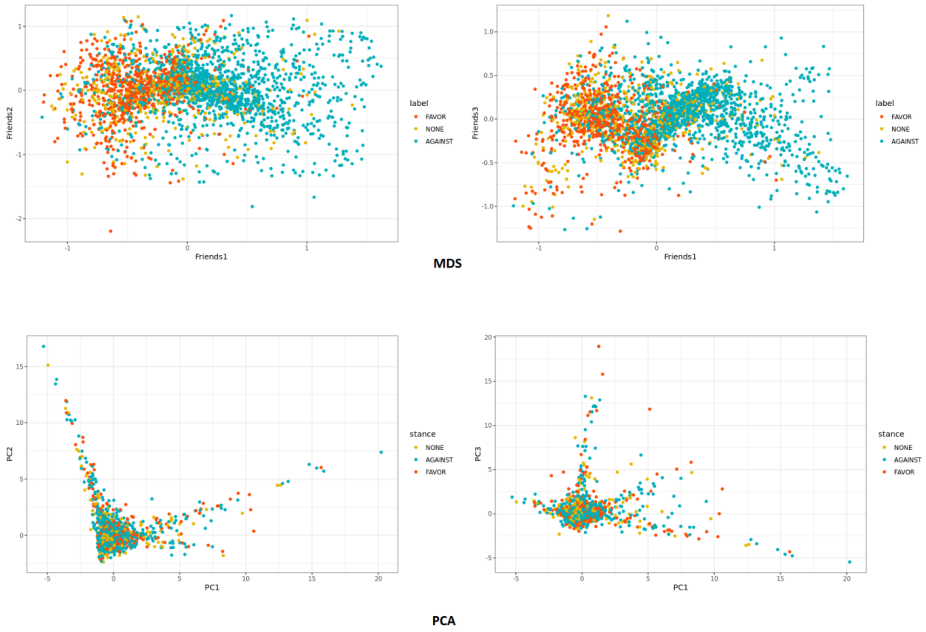


Figure 1. Scatter plots of the First-Second dimensions and First-Third dimensions extracted by the MDS from the distance matrix of the Friends Network and by the PCA of the DTM.

3. Analysis

3.1. Experiments

The SardiStance competition included two tasks: in task A participants had to predict the stance towards the movement Sardine exploiting only textual information, while in task B teams could employ contextual information based on: the post, the user, and his/her social networks. The SardiStance organizers then provided a training set (N=2132), a test set (N=1110) and separate files containing information on tweets and users, as well as edge lists for the different relationships between users. Predictive systems were evaluated using the F1-score computed over the two main classes (favor and against). For each task, organizers also computed a baseline using a simple statistical learning model based on SVM combining uni-gram features.

In this paper we show the results of three experiments: the first experiment concerns the prediction of stances using only the first 100 principal components of the DTM (task A), the second adds to the features the 8 dimensions derived from the MDS (task B), the third also adds some context information (task B), such as the number of: posts, friends, and followers for each user; listings, retweets, and favorites for each tweet.

We chose as statistical learning model a penalized multinomial logistic regression, using a 5-fold cross-validation (with 100 repetitions) tuning of the decay parameter through the `nnet` R package. Therefore, we used a single-hidden-layer neural network, where the aim of the decay parameter is to prevent the weights to be too large through penalizations during the fitting process. This procedure should help avoiding overfitting (Venables et al., 2013). We chose this classifier because, even though it does not need the independent variables to be statistically independent from each other (unlike other classifiers, e.g., Naïve Bayes), it assumes collinearity to be low (Belsey, 1991).

3.2. Results

Table 1 shows for each experiment the global level of accuracy, the F1-score for the stances 'against' and 'favor', and their F1-average. Measures were computed on the labeled (gold) test set. The last column shows the ranking position in which each model would have been placed if it had participated in the relative task of the competition. The results of the application of these techniques show levels of accuracy comparable to classical feature extraction techniques and to the best task models, despite using potentially a much smaller number of predictors, and without considering n-grams. Moreover, our models for task B did not include other information on the tweet itself (source, creation date, etc.), contextual information about the user (creation date, emoji in bio, etc.), quotes and replies networks.

Table 1. Results of the experiments (test set – gold labels) using 100 principal components of the DTM (PC), 8 dimensions extracted from MDS coordinates (MDS), and contextual information (context).

Experiment	Accuracy	F1-against	F1-favor	F1-Avg (against, favor)	Rank in task[3]
1 (PC)	0.6099	0.7506	0.3325	0.5415	16th
2 (PC,MDS)	0.7189	0.8440	0.5950	0.7195	5th
3 (PC,MDS,context)	0.7180	0.8426	0.6095	0.7261	3rd

In the SardiStance competition, our team, TextWiller (Ferraccioli et al., 2020), also used 50 vectors derived from word embedding as features, improving F1-average of Experiment 3 by a very small amount (+0.0048). This result does not particularly surprise us, because we have already found that using word

[3] www.di.unito.it/~tutreeb/sardistance-evalita2020/index.html

embeddings as predictive features does not improve the results compared to frequencies (Sciandra, 2020) or TF-IDFs of words, as in this work. We believe that a smarter use of word embeddings, e.g., by means of the dependency relations between the target of interest (Sardines here) and the connected tokens in the dependency tree (Lai et al., 2021), can lead to better results. Instead, the inclusion of information about social networks increased the F1-average remarkably (+0.18). In fact, the Experiment 1 with only the principal components of the DTM ranks below the baseline of about 0.037 points of the F1-average, whereas Experiment 2 ranks 0.09 points above the baseline of Task B. Moreover, the analysis of the most important variables[4] in Experiment 3 also shows the dominance of network measures (Fig. 2).

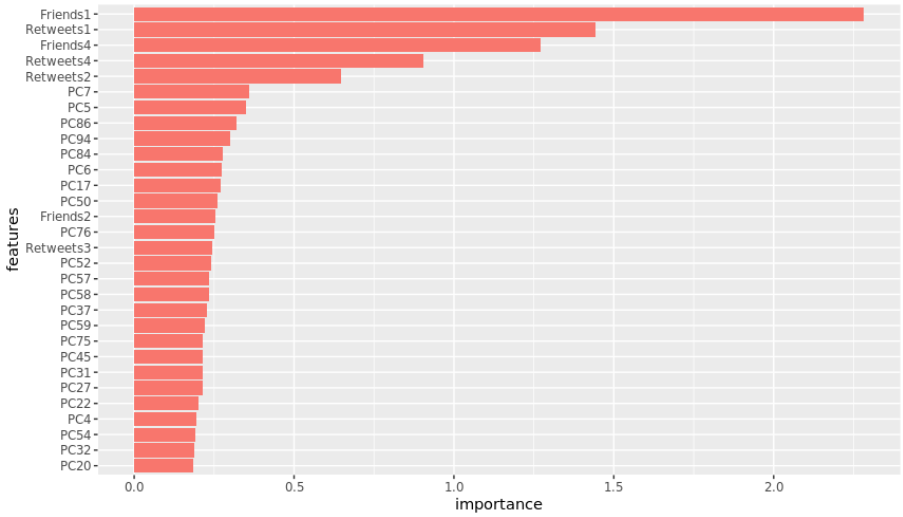


Figure 2. Variable importance (top 30) – Experiment 3

Specific tests on each variable, such as Wald's test or Likelihood Ratio test, may be useful to complete the analysis, along with the relative risks (odds ratios).

4. Conclusion

The main advantage of this approach, in our opinion, lies in the possibility to extract a reduced number of orthogonal dimensions that maximize the ability to capture the variability of textual features. The approach of this work proved to

[4] The method of computing the importance of variables in a neural network exploits combinations of the absolute values of the weights (Gevrey et al., 2013).

be also highly competitive with respect to dimensionality reduction techniques related to the semantic scope, such as word embedding. Clearly, we need to improve the performance of models based only on PCA. For instance, more elaborate state of the art methods of feature extraction for texts (e.g., n-grams, stylistic, lexicon-based, and dependency-based features) surely can enhance the prediction accuracy over our approach, but they often result in very sparse and/or highly collinear matrices, leading to possible computational problems and overfitting risk.

Finally, we point out that reducing the dimensionality of the ties among users showed great efficiency, significantly improving the performance of the prediction model with an extremely limited number of dimensions.

Acknowledgements

This research has been partially supported by the University fund for Research (FAR 2020) of the University of Modena and Reggio Emilia.

References

Belsley D. (1991). *Conditioning diagnostics: collinearity and weak data in regression*. New York: Wiley.

Cignarella A. T., Lai M., Bosco C., Patti V. and Rosso P. (2020). SardiStance@ EVALITA2020: Overview of the task on stance detection in Italian tweets. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (pp. 1-10). CEUR Workshop Proceedings.

Ferraccioli F., Sciandra A., Da Pont M., Girardi P., Solari D. and Finos L. (2020). TextWiller@ SardiStance, HaSpeede2: Text or Con-text? A smart use of social network data in predicting polarization. *Proc. of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020).* CEUR Workshop Proceedings.

Gevrey M., Dimopoulos I. and Lek S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3): 249-264.

Klimenta M. and Brandes U. (2012). Graph drawing by classical multidimensional scaling: new perspectives. In *International Symposium on Graph Drawing*. Springer, Berlin, Heidelberg, pp. 55-66.

Küçük D. and Can F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, *53*(1): 1-37.

Lai M., Cignarella A. T., Finos L. and Sciandra A. (2021). WordUp! at VaxxStance 2021: Combining Contextual Information with Textual and Dependency-Based Syntactic Features for Stance Detection. In *XXXVII International Conference of the Spanish Society for Natural Language Processing.* (Vol. 2943). CEUR Workshop Proceedings, pp. 210-232.

Mohammad S. M., Sobhani P. and Kiritchenko S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, *17*(3): 1-23.

Sciandra A. (2020). COVID-19 Outbreak through Tweeters' Words: Monitoring Italian

Social Media Communication about COVID-19 with Text Mining and Word Embeddings. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, pp. 1-6.
Solari D., Sciandra A. and Finos L. (2019). TextWiller: Collection of functions for text mining, specially devoted to the Italian language. *Journal of Open Source Softwar*e, 4(41): 1256.
Venables W. N. and Ripley B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

A two-stage LDA algorithm for ranking induced topic readability

Mariangela Sciandra, Alessandro Albano

University of Palermo – mariangela.sciandra@unipa.it
alessandro.albano@unipa.it

Abstract
Probabilistic topic models, such as LDA, are standard text analysis algorithms that provide predictive and latent topic representation for a corpus. However, due to the unsupervised training process, it is difficult to verify the assumption that the latent space discovered by these models is generally meaningful and valuable.
This paper introduces a two-stage LDA algorithm to estimate latent topics in text documents and use readability scores to link the identified topics to a linguistically motivated latent structure. We define a new interpretative tool called *induced topic readability*, which is used to rank topics from the one with the most complex linguistic structure to the one with the lowest semantic content readily. The usefulness of our method is shown with an application to real data, using articles from the New York Times.

Keywords: topic models; Latent Dirichlet Allocation; readability; coherence

1. Introduction

Probabilistic topic models are one of the most widespread machine learning techniques for textual analysis, especially since the introduction of the Latent Dirichlet Allocation (LDA) Blei et al. (2003). These models build latent topics as multinomial distributions over words and assume that each document can be described as a mixture of these topics. Once the models are trained, they provide a framework for humans to understand document collections directly by "reading" models or indirectly by using topics as input variables for further analysis Boyd-Graber et al. (2017). Often, in qualitative studies, the goal is to find meaningful and interpretable topics. Automated estimates of topics coherence rely on the frequency of word co-occurrences in a reference corpus Lau et al. (2014); Newman et al. (2010); Aletras and Stevenson (2013); Ramrakhiyani et al. (2017).
Although much research has been devoted to the semantic coherence of topics, no effort has yet been made to investigate the relationship between topics and text readability Kincaid et al. (1975). Readability Kincaid et al. (1975) measures how easy it is to read and understand a text, depending on its unique linguistic features. Moreover, readability also influences how clearly a reader can

understand a text.
Therefore, the research question is: could the identified topics be associated with different readability levels of each? The main problem in answering this question comes from the difficulty in measuring the readability of topics since topics are multinomial distributions over words. This paper defines a two-stage LDA algorithm for measuring a new concept called induced topic readability that should help understand the detected topics and rank them according to their linguistic complexity. Section 2 is devoted to a brief recap of the standard LDA procedure; in Section 3, the most used readability indices are described. Section 4 introduces our proposal and includes an application to real data. Conclusions end the paper.

2. Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over an unknown number (fixed *a priori*) of latent topics, where a distribution of words Blei et al. (2003) characterizes each topic. Latent Dirichlet allocation posits a "generative process" about how the data came to be Boyd-Graber et al. (2017). In the model, the words of the documents are the observed variables, whereas the topic structures are the hidden variables. The problem of inferring the hidden topic structure from the documents consists in computing the poste- rior distribution of topic structures, that is, the conditional distribution of the hidden variables given the documents Blei (2012). These 'hidden' topics are then surfaced based on the likelihood of word co-occurrence. Formally, this is a Bayesian Inference problem with the main drawback that consists in identifying the correct number of topics present in the corpus (Arun et al., 2010).
Let z be a topic, w a collection of words. A generic document d is assumed to be a mixture of topics. The LDA aims to understand how many and which topics are present in the corpus documents and how strong their presence is. Specifically, for each topic z ($z = 1, \ldots, Z$), ϕ_z denotes the probability distribution over the words of the observed vocabulary. In other words, ϕ_z describes the probability for each word to appear in a given topic. Then, each document d ($d = 1, \ldots, D$) has a probability distribution over topics, θ_d, which encodes the probability for each topics to be drawn in a given document:

$$p(w|d) = p(w|z)p(z|d) \quad \forall \quad d = 1, \ldots, D, \qquad z = 1, \ldots, Z. \tag{1}$$

3. Readability

A readability index (R) estimates how difficult a text is to read. Its value is related to a text's measure of complexity from some quantitative attributes of texts such as word lengths, sentence lengths, syllable counts, etc... Text complexity is

then compared to how well readers comprehend the text. In literature, several readability indices have been proposed (Flory et al., 1992): they all share the core idea of interpreting readability as a measure of text complexity but differ in how they emphasize particular aspects of text complexity. Some emphasize syllable counts, while others look only at the word and sentence lengths.

· Automated Readability Index (ARI):

$$ARI = 4.71 \cdot (characters/words) \tag{2}$$

It is derived from ratios representing word difficulty (number of characters per word) and sentence difficulty (number of words per sentence)

· Flesch Reading Ease (FRE):

$$FRE = 206.835 - 1.015 \cdot (Total\ Words/Total\ Sentences) - 84.6 \cdot (TotalSyllables/TotalWords) \tag{3}$$

Higher scores indicate material that is easier to read; lower numbers mark harder-to-read passages.

· Rate Index (RIX):

$$RIX = (LongWords/Sentences) \tag{4}$$

where long words are words where number of characters is greater than 6. RIX assumes scores between 0 (very easy) and 55 (very difficult).

A complete comparison between readability formula based on psycholinguistic and cognitive models of reading, and traditional readability formulas on a large corpus of texts can be found in Crossley et al. (2011).

4. Proposal

The sketch of our algorithm can be summarized as follows:

1. Estimate an LDA model, and extract the topic proportions θ_d
2. Derive the conditional document probabilities, given a topic, via Bayes rule:

$$P(d|\phi_z) = \frac{P(\phi_z|d)P(d)}{P(\phi_z)}, \tag{5}$$

Where

$$P(\phi_z|d) = \theta_d,\ \ P(\phi_z) = \frac{1}{D}\sum_{d=1}^{D} P(\phi_z|d),\ \ P(d) = \frac{1}{D} \tag{6}$$

3. Compute an automated readability index for each document of the corpus $R_1, \ldots, R_d, \ldots, R_D$.
4. For each topic $z = 1, \ldots, Z$ compute the *induced readability index* as a convex combination:

$$IR_z = \sum_{d=1}^{D} P(d|\phi_z) R_d \tag{7}$$

5. Produce the final ranked list of topics, by sorting them in decreasing order according to t *induced readability* value.

Note that the range of IR_z depends on which readability index R_d is chosen for analysis. In fact, the choice of Rd is not unambiguous but must be based on the type of document (newspaper articles, poems, songs, etc.) and on its linguistic features.

4.1. Real data analysis

The proposed algorithm is applied to a dataset of 1 000 articles that appeared in the *New York Times* between April and July 2016[1]. Firstly, the readability of each article of the dataset is calculated using the Flesh Reading Ease score (Flesch, 1948). Then, data preprocessing is performed, and the LDA model is trained in R setting the number of topics is equal to 50.

For each topic z, we calculate the proposed *induced readability index*, IR_z and rank the topics according to these values. In order to evaluate the goodness of the detected topic structure, topics are also ranked by using a measure of coherence following Simonetti et al. (2022) indicating how semantically linked the most probable words in a topic are. Note that the two measures have the same codomain, between 0 and 100. The scatterplot in Fig. 1 shows the coherence and readability for the estimated 50 topics. Each dot represents a single topic; the dashed lines indicate the average respectively of readability and coherence indices and divide the plane into four quadrants.

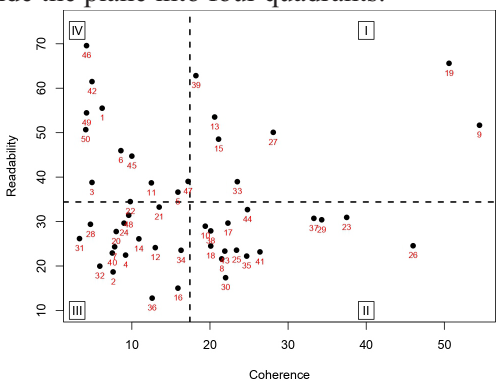


Figure 1. Coherence vs Readability

[1] https://www.kaggle.com/nzalake52/new-york-times-articles.

As expected, the two measures are uncorrelated; the Pearson correlation coefficient is 0.05. In fact, these indices measure two different aspects of the quality of a topic. For example, a topic can be strongly coherent but be associated with texts with low readability scores, thus belonging to the II quadrant (e.g., topic 26). On the other hand, topics belonging to the I quadrant are highly coherent since they identify well-defined themes and, at the same time, are associated with high readable articles. For example, topic 9, whose most probable words are: "Donald Trump, Republican, Hillary Clinton, Campaign, Party, Voter, Democratic, Presidential, Bernie Sanders, Candidate". This ex- ample allows us to clarify how the aspect related to the linguistic structure is something different from the coherence in the evaluation of the meaning of the topics identified. It is impossible to define which of the two orderings should be preferred; the choice will depend on the scope of our analysis.

5. Conclusions

The two-stage LDA algorithm jointly estimates topics, ϕ_z, and topic *induced readability* scores, IR_z. This method has important implications concerning the evaluation of a topic model. Indeed, the first ranked topic characterizes the most readable documents, thus can be considered the easiest to go into. We also stress that our topic-readability measure can be used in conjunction with other semantic coherence measures. Readability and coherence can be considered "complementary". An interested user should use both pieces of information to explore better the latent semantic space estimated by a topic model.

References

Aletras N. and Stevenson M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Arun R., Suresh V., Madhavan C. V. and Murthy M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer.

Blei D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boyd-Graber J. L., Hu Y. and Mimno D. (2017). Applications of topic models (Vol. 11). now Publishers Incorporated.

Crossley S. A., Allen D. and McNamara D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23:84– 101.

Flesch R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Flory S. M., Phillips T. J. and Tassin M. F. (1992). Measuring readability: A comparison of accounting textbooks. *Journal of Accounting Education*, 10(1):151–161.

Kincaid J. P., Fishburne Jr R. P., Rogers R. L. and Chissom B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Lau J. H., Newman D. and Baldwin T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Newman D., Lau J. H., Grieser K. and Baldwin T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Ramrakhiyani N., Pawar S., Hingmire S. and Palshikar G. (2017). Measuring topic coherence through optimal word buckets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 437–442.

Simonetti A., Albano A., Plaia A. and Tumminello M. (2022). Statistically validated networks for evaluating coherence in topic models. In *The 10th International Conference on Complex Networks and their Applications*.

Text mining for socially responsible behavior: fostering the adoption of green technologies

Ludovica Segneri[1], Andrea Fronzetti Colladon[1,2], Claudia Fabiani[1,3], Anna Laura Pisello[1,3]

[1]University of Perugia – ludovica.segneri@collaboratori.unipg.it andrea.fronzetticolladon@unipg.it caludia.fabiani@unipg.it anna.pisello@unipg.it
[2]Kozminski University
[3]CIRIAF - Interuniversity Research Centre

Abstract
The global population is facing significant socio-political, economic, and environmental challenges, in part due to periodic crises occurring around the world, caused by natural events and anthropogenic actions. As a consequence, people are exposed to different information concerning, for instance, climate change, global warming, technological advances and their impact on society.
The Internet and the availability of online high-quality data sources present a real opportunity to understand how people are being made aware of new energy solutions and their advantages for environmental sustainability and energy conservation. In this scenario, analyzing online news, social media platforms like Twitter, or search volume data (like Google Trends) becomes particularly relevant. Here, we focus on online news and consider media communication in different European countries, to understand their impact on citizens' behavior and plan future interventions.
In particular, we assess how media shape the discourse around green technologies and how this relates to differences in citizen behavior. We combine methods and tools of text mining and social network analysis (Fronzetti Colladon et al., 2020) to assess the semantic importance of Energy-Related-Terms (ERTs) using the Semantic Brand Score (SBS) indicator (Fronzetti Colladon, 2018; Fronzetti Colladon & Grippa, 2020). The SBS is made of three dimensions: (i) prevalence, (ii) diversity, and (iii) connectivity. Prevalence identifies the frequency of appearance of ERTs in a discourse and evaluates their visibility. Diversity measures the heterogeneity and uniqueness of an ERT's image. Connectivity represents the brokerage power of ERTs, i.e. their ability to link different discourse topics.
Results show trends in semantic importance of ERTs and their conveyed image. Among ERTs, we look at, for example, renewables, storage technologies, and energy communities. We propose strategies to improve the awareness, memorability and understanding of these discourse themes – with the ultimate goal of broader

implementation and adoption of green technologies.
The SBS approach is applicable in a variety of energy-related settings, potentially with the aim of supporting the diffusion of new standards, policies and energy saving practices. The same methodology can also be applied to identify information gaps, which may be reduced through improved media coverage.
Keywords: green technologies; societal awareness; online news; semantic brand score; semantic network analysis; energy transition; text mining.

Funding
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No890345.

References
Fronzetti Colladon, A. (2018). The Semantic Brand Score. *Journal of Business Research*, *88*, 150–160. https://doi.org/10.1016/j.jbusres.2018.03.026
Fronzetti Colladon, A., Gloor, P., & Iezzi, D. F. (2020). Editorial introduction: The power of words and networks. *International Journal of Information Management*, *51*, 102031. https://doi.org/10.1016/j.ijinfomgt.2019.10.016
Fronzetti Colladon, A., & Grippa, F. (2020). Brand Intelligence Analytics. In A. Przegalinska, F. Grippa, & P. A. Gloor (Eds.), *Digital Transformation of Collaboration* (pp. 125–141). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-48993-9_10

Exploration de contextes définitoires en français : cas d'étude sur la COVID-19

Gerardo Sierra[1,3], Juan-Manuel Torres-Moreno[2], Mercè Lorente[3]

[1]Universidad Nacional Autónoma de México – gsierram@iingen.unam.mx
[2]Université d'Avignon – juan-manuel.torres@univ-avignon.fr
[3]Universitat Pompeu Fabra – merce.lorente@upf.edu

Abstract
Within different knowledge bases, computational linguistics shows a growing interest in the automatic identification of definitional contexts (DC) found in specialised texts. These are understood as the textual fragments containing a term and its definition, usually linked by definitional patterns, such as certain verbs, typographical marks and discursive markers. Recognition of regular structures of DC allows the automatic retrieval of these contexts. This paper presents an exploration of the regular structures of DC in French, applied to a corpus on COVID-19. First of all, a methodology of the main constituents of the DC as well as the typology used on terms and definitions is proposed. Thus, the search for the most well-known verbal definitional patterns and discourse markers in the literature is carried out. From the obtained fragments, XML tags are used to tag the components of each DC. Thus, a grammar of the DC in French is obtained and their productivity is evaluated. The DC obtained for the COVID-19 corpus are also exemplified.

Keywords: Conceptual information extraction, Definitional context, Corpus, Pattern recognition, Terminology, COVID-19

Résumé
Au sein de différentes bases de connaissances, la linguistique computationnelle s'intéresse de plus en plus à l'identification automatique des contextes définitoires (CD). Ces contextes, trouvés dans les textes spécialisés, comme des fragments textuels où un terme et sa définition sont introduits, généralement liés par des motifs de définition, tels que certains verbes, des marques typographiques et des marqueurs de discours. La reconnaissance des structures régulières des CD permet de les retrouver automatiquement. Cet article présente une exploration des structures régulières des CD en français, appliquée à un corpus sur la COVID-19. Tout d'abord, une méthodologie des principales composantes des CD et de la typologie utilisée sur les termes et définitions est proposée. C'est ainsi que commence la recherche des schémas verbaux

définitoires et des marqueurs de discours les plus connus dans la littérature. À partir des fragments obtenus, les composants de chaque CD sont étiquetés à l'aide de balises XML. Ainsi, une grammaire des CD en français est obtenue et leur productivité est évaluée. Elle est également illustrée par les CD obtenus pour le corpus COVID-19.

Mots clés: Extraction d'information conceptuelle, Contexte définitoire, Corpus, Reconnaissance de patrons, Terminologie, COVID-19

1. Introduction

L'acquisition, la récupération, la construction et le développement des systèmes basés sur les connaissances est l'un des objectifs du domaine de l'ingénierie des connaissances. En particulier, la linguistique informatique cherche à extraire automatiquement des informations à partir de données non structurées afin de créer des bases de connaissances. Dans le domaine de l'extraction terminologique, l'organisation et la description des concepts présents dans les textes spécialisés suscitent un intérêt croissant. Cela constitue la base pour la construction d'ontologies représentant le réseau conceptuel d'un domaine spécifique (Acosta et Aguilar, 2021), l'identification d'unités phraséologiques spécialisées (Gallego-Hernández, 2022), la création de bases de données de connaissances lexicales et l'élaboration de glossaires ou de dictionnaires spécialisés (Mirapeix et al., 2020). Elles présentent également un intérêt pour les outils de question-réponse (Saggion et Gaizauskas, 2004) et les applications d'apprentissage en ligne (Westerhout et Monachesi, 2007). Un certain nombre de recherches ont été faites en vue de la reconnaissance et de la classification des termes et des définitions en tant que domaine théorique et appliqué. Entre autres, on peut citer les premiers travaux d'Auger (1997) et de Rebeyrolle et Tanguy (2000) sur les verbes utilisés dans les énoncés définitoires et les travaux de Meyer (2001) sur la recherche systématique de contextes riches en connaissances, et les développements de systèmes d'extraction automatique de contextes définitoires pour l'espagnol (Alarcón et al., 2007) et pour le catalan (Espinosa-Anke et Saggion, 2014). Il faut également souligner les corpus de contextes définitoires pour différentes langues : en allemand, avec 174 éléments (Storrer et Wellinghoff, 2006) ; en anglais, avec 1 700 éléments (Navigli et al, 2010) ; et en espagnol, avec 1 424 éléments (Sierra et al., 2006). L'extraction automatique de contextes définitoires consiste à identifier les phrases où un terme et sa définition sont introduits, liés par un motif définitoire. Ce motif peut être un patron typographique, un verbe définitoire ou un marqueur de reformulation. L'identification de la structure des contextes définitoires, c'est-à-dire des modèles de comportement syntaxique et sémantique, facilite l'obtention de données pour la recherche et la classification des définitions. Cette identification facilite également leur extraction automatique car elle permet d'identifier leurs éléments et l'ordre dans lequel ils apparaissent. Cet article contribue à l'exploration des

éléments constitutifs des contextes définitoires en français, basé sur des travaux d'Auger et Rebeyrolle, appliqués à un corpus traitant le sujet de la COVID-19. D'une part, l'objectif est de constituer un corpus de contextes définitoires dûment étiqueté avec leurs éléments constitutifs., ce qui va permettre de proposer leur structure afin de faciliter leur extraction automatique. D'autre part, les contextes définitoires obtenus seront analysés afin de fournir des lignes directrices sur les termes et concepts associés à la COVID-19, enrichissant ainsi les bases de données terminologiques actuelles et les connaissances sur le domaine. Il est à noter que le corpus et les résultats obtenus seront disponibles en ligne.

2. Travaux précédents

Auger (1997) explore les prédications qui introduisent et déterminent l'information conceptuelle associée à une définition, les énoncés définitoires, ainsi que leurs éléments constitutifs : le terme à définir, l'expression définitoire et la particule copulative les associant. Sur la base de cette particule copulative, Auger distingue les énoncés définitoires métalinguistiques (« désigner », « être le nom de », « s'appeler ») des énoncés définitoires linguistiques (« être », « autrement dit », « comprendre »). Cet auteur analyse et évalue la productivité des modèles syntaxiques réguliers pour les différents types d'énoncés définitoires dans des corpus de textes spécialisés. Rebeyrolle et Tanguy (2000) décrivent une méthodologie pour l'extraction d'énoncés définitoires à partir de modèles morphosyntaxiques dans un corpus de textes spécialisés. Ils établissent une classification entre les énoncés définitoires : de désignation (« désigner »), de dénomination (« s'appeler »), de signification (« signifier »), introduits par « c'est-à-dire », de classification (« est un ») et parenthétiques. Ils extraient les modèles qui contient certains types de verbes ou de structures verbales. Leur évaluation a montré la qualité de certains modèles, tandis que d'autres ont tendance à récupérer une grande quantité de bruit. Malaisé (2005) extrait des définitions non spécialisées en français pour la constitution d'ontologies. Les définitions sont introduites à partir de verbes métalinguistiques (dénommer, définir), de marqueurs métalinguistiques (définition, nom) en combinaison avec des verbes (référer, utiliser), de marqueurs de discours (c'est-à-dire, en d'autres termes) et de signes de ponctuation, spécifiquement des parenthèses. Il a constaté que plus les modèles sont nombreux, plus la probabilité que le contexte soit un énoncé définitoire est grande. Au sein du Groupe d'Ingénierie Linguistique (http://www.iling.unam.mx), un projet a été développé dans le but d'extraire automatiquement des contextes définitoires (CD) (Sierra, 2009). La base de l'extraction réside dans les modèles définitoires. Ainsi, la définition, la typologie et le rôle joué par les prédications verbales dans les types de définition ont été analysés. La méthodologie de l'Extracteur de contextes définitoires a été proposée (Alarcón, 2009). Cet extracteur prend en entrée une liste de modèles

de verbes définitoires et en sortie les CD classés par types de définition, comme indiqué à la section 4.

3. Contextes définitoires

Un élément clé de la reconnaissance automatique de CD est l'identification des modèles utilisés pour lier un terme à sa définition ou pour souligner visuellement sa présence dans un texte (Estopà, 2001). Le prototype de CD est une structure discursive composée de deux éléments minimaux : un terme (T) et sa définition (D), reliés l'un à l'autre par un modèle définitoire (DP). Parmi les modèles définitoires, nous trouvons deux classes générales : les typographiques et celles syntaxiques. Dans ces dernières, et selon les éléments qui y sont présents, on peut trouver de modèles verbaux et/ou de marqueurs de reformulation. Ces marqueurs constituent un processus de reformulation dans lequel le sens d'un terme est expliqué sur la base de structures syntaxiques non verbales. Dans le cas de CD, elles servent à désigner les termes comme des éléments de la langue elle-même (Bach, 2005). En outre, d'autres informations métalinguistiques et pragmatiques peuvent être présentes concernant la forme, les conditions d'utilisation ou la portée opérationnelle des termes. De telles informations correspondent à ce que nous appelons un schéma pragmatique (PPR). Parmi les constructions verbales syntaxiques qui relient un terme à sa définition, on trouve les verbes métalinguistiques, c'est-à-dire ceux utilisés pour faire référence à la langue elle-même, comme les verbes « définir », « comprendre » ou « dénommer ». On trouve également des verbes de langue générale très courants, utilisés dans différentes situations communicatives et non seulement dans des situations définitoires, comme « être » et « considérer ». Une caractéristique de certains verbes est d'apparaître avec un connecteur (la préposition « par » et l'adverbe « comme ») et dans d'autres, il est courant d'utiliser des particules grammaticales telles que le pronom personnel on le réflexif « se » (« on entend », « se définit comme »).

4. Méthodologie

4.1. Recherche de modèles définitoires

Un corpus de documents en français a été crée à l'aide de la plateforme en ligne Sketch Engine[1]. À cette fin, 20 termes sélectionnés de Termium Plus[2] ont été utilisés. Au total, un corpus de 351 documents contenant 2 145 318 mots a été obtenu. Ce corpus est appelé COVID. À l'aide de l'outil Concordances de Sketch

[1] https://www.sketchengine.eu/

[2] Termium Plus est le Banque terminologique et linguistique du Gouvernement du Canada, https://www.btb.termiumplus.gc.ca/publications/covid19-fra.html

Engine, divers modèles définitoires ont été recherchés manuellement dans le corpus COVID. Une classification a été faite entre ceux énoncés qui sont des contextes définitoires (CD) et ceux qui ne le sont pas (Non CD). Le tableau 1 montre le nombre total d'occurrences pour chaque type en fonction du modèle recherché.

Tableau 1. Occurrences des modèles définitoires dans le corpus COVID.

Modèles définitoires	CD	Non CD
[lemma="définir"][]{0,6}[word="comme"]	35	3
[lemma="définir"][]{0,6}[word="par"]	4	13
[lemma="entendre"][]{0,5}[word="par"]	5	3
[word="par"][]{0,5}[lemma="entendre"]	2	0
[lemma="désigner"]	32	31
[lemma="être"][tag="D.*"tag="Z.*"]	864	1352
[word="le"][word="mot"|word="nom"|word="terme"][]{1,6}[lemma="être"]	12	4
autrement dit	4	4

4.2. Étiquetage du corpus

Une fois sélectionnés les contextes définitoires, l'étiquetage de leurs éléments constitutifs a été effectué. Pour cela, des balises XML sont utilisées comme base pour identifier les parties d'un CD. L'exemple suivant montre la façon dont l'étiquetage a été réalisé :

Les seconds tests disponibles sont les tests dits « antigéniques » : ils détectent la présence de protéines de structure du virus (lesquelles sont des « antigènes », un terme qui désigne tout élément étranger à l'organisme capable de déclencher une réponse immunitaire), avec une bonne sensibilité durant la période se situant autour du pic de charge virale.

```
<CD num="1">Les seconds tests disponibles sont les <T num="1a"> tests </T> <MRD> dits
</MRD> <PT>« </PT> <T num="1b"> antigéniques </T> <PT>» </PT> <PT>:</PT> <D num="1">
ils détectent la présence de protéines de structure du virus</D>(lesquelles sont des
<PT>«</PT> <T num="2"> antigènes </T><PT>» </PT>,<PVD> <MRD>un terme qui</MRD><VD
lemme="désigner"> désigne </VD> </PVD><D num="2"> tout élément étranger à l'organisme
capable de déclencher une réponse immunitaire </D>), avec une bonne sensibilité durant
la période se situant autour du pic de charge virale.<p/> </CD>
```

CD : Contexte définitoire. Une phrase ne peut avoir qu'un seul CD, bien que celui-ci puisse être complexe, comme dans l'exemple ci-dessus. Nous avons un terme qui est introduit par un indice typographique définitoire (*« »* :) et un autre introduit par le PVD, composé d'un MRD et d'un VD (un terme qui désigne).
T : Terme. Dans un CD, il y a au moins un terme. Dans l'exemple ci-dessus, il y a deux termes. Le premier (tests antigéniques) est coupé par un MRD, il est donc composé de deux parties.

D : Définition. Il doit y exister au moins une définition. Dans l'exemple, il y a une définition pour chacun des deux termes, mais ce n'est pas nécessaire que le nombre de termes et de définitions sur un CD soie le même.
PVD : Prédicat verbal définitoire. Il est formé par (MARD) + VD + (NX).
VD : Verbe définitoire. Il est un syntagme verbal qui relie le terme à la définition. Le lemme du verbe principal est indiqué.
NX : Connecteur. Il indique la fonction du VD. Les principaux sont comme, par et les déterminants.
PP : Modèle pragmatique. Il peut être contextuel, d'auteur ou temporel. Deux ou trois types de modèles pragmatiques peuvent même apparaître l'un après l'autre. Il peut se trouver avant le terme ou entre le verbe définitoire et le connecteur.
MRD : Marqueur de reformulation définitoire. Parmi d'autres trouvés dans le corpus : est un terme souvent utilisé, sous le terme, un terme qui, connu sous le nom de, autrement dit.
PT : Modèle typographique. Les éléments qui renforcent un élément du CD, tel que le terme ou la définition, ou qui remplacent ou complètent le PVD, sont notés. Il peut s'agir d'un signe de ponctuation, d'une parenthèse ou d'une fin de ligne.

5. Résultats[3]

5.1. Structures syntaxiques des CD

La structure générale la plus courante, indépendamment des modèles typographiques, est la suivante :
(PP) T PD D
où PD est VD + NX. C'est-à-dire, d'abord le terme, puis le modèle définitoire et enfin la définition. Le modèle définitoire vient immédiatement après le terme, mais il peut aussi y avoir un séparateur (virgules ou points) ou un signe de regroupement (parenthèses ou guillemets). Les structures trouvées sont disponibles en ligne. Le tableau 2 montre quelques exemples des structures trouvées.

Tableau 2. Exemples de structures syntaxiques des contextes définitoires dans le corpus COVID

T VD PP NX D	T MRD NX VD D	T, VD NX D1 (ou D2)
PP VD T NX D	T s/ MRD VD D	T VD NX T1 D1;T2 D2;T3 D3;T4 D4
MD T VD PP NX D	T s/ T VD NX D	T (MRD VD D)

[3] Le corpus et les résultats complets sont disponibles à l'adresse électronique : github.com/GIL-UNAM/contextes_definitoires.

5.2. Analyse des modèles verbaux

A partir des CD obtenus (tableau 2), on observe les points suivants.

définir comme. Le schéma verbal *définir comme* est un des schémas les plus productifs, avec une précision de 92,10%, résultats très cohérents avec ceux obtenus par Rebeyrolle et Tanguy (92,59%).

définir par. Contrairement à *définir comme*, ce modèle n'est pas aussi productif, puisque seulement 6 contextes définitoires et 11 contextes non définitoires ont été obtenus.

entendre par. Deux formes ont été recherchées : « *entendre par »* et « *par entendre »*, avec une distance de 5 mots. Sous l'une ou l'autre forme, elles sont utilisées avec la particule *on*. De la première forme, sa structure est classique : T VD NX D. De la deuxième forme, sa structure est la suivante : « *par* T *on entend* D ».

désigner. Pour « *désigner »*, plusieurs marqueurs de reformulation ont été trouvés en comparaison avec les autres verbes définitoires.

être. Le verbe « *être »* est un connecteur prototypique reliant un sujet à son prédicat, sa fréquence d'utilisation est donc très élevée. Il a été considéré comme le modèle définitoire le plus fréquemment utilisé, au même temps que le déterminant *un*. En anglais, il est connu sous le nom de ISA. Bien qu'il soit le plus utilisé, cela ne signifie pas qu'il soit le plus productif.

le mot|nom|terme est. Ce modèle, reconnu par Auger, a pour variante l'omission du déterminant *un*. La plupart des définitions obtenues dans les CD sont synonymiques (celles qui indiquent l'équivalence entre deux termes). Dans certains cas, une brève définition est donnée, qui peut également faire office de terme.

5.3. Lexique du COVID

L'analyse des contextes définitoires nous permet de constater que les termes peuvent aller au-delà de ce qui est normalement défini dans le domaine de la terminologie. Les termes peuvent même être des noms propres, c'est-à-dire des personnes et des pays. Mais ce sont des termes précisément parce qu'ils sont associés à un concept, c'est-à-dire qu'il y a quelque chose qui les définit. Du point de vue de la lexicographie, les pays sont définis par le genre proche ou *genus* (pays ou lieu) et la différence spécifique, qui dans ce cas est la caractéristique qui les rend pertinents par rapport à la pandémie. Du point de vue de la terminologie (Cabré, 1999), les unités lexicales activent leur valeur terminologique en contexte dans un discours de spécialité, qui dans ce cas concerne la COVID. Un exemple des termes obtenus et de leurs définitions est donné dans le tableau 3.

Tableau 3. Exemples de termes et de définitions obtenus à partir du corpus COVID.

Terme	**Définition**
<T>période d'exposition au coronavirus</T>	<D num="1">la période d'incubation avant l'apparition des symptômes</D> <D num="2">ou de la date de notification du cas lorsque les données concernant le début des symptômes n'était pas disponible</D>
<T>covidés</T>	<D>ceux qui ont contracté la maladie à coronavirus, un peu comme on dit les grippés, les enrhumés</D>
<T num="1">confinement</T> <T num="2">circuit breaker</T>	<D>casser le cycle de la transmission du virus</D>

En comparant la terminologie fournie par Termium Plus et le Sketch Engine avec celle obtenue à partir des contextes définitoires, on observe une augmentation substantielle des termes. Les définitions, en revanche, ne sont pas nécessairement des définitions standard. Bien qu'elles puissent être analytiques, elles peuvent être incomplètes. En effet, les textes de spécialité ne définissent pas les termes utilisés dans un seul énoncé, mais tout au long du texte.

6. Conclusions

L'identification de modèles définitoires, verbaux et typographiques, est un élément qui s'est avéré efficace pour obtenir une base de connaissances des termes et des définitions dans un domaine de spécialité. Dans le cas du corpus COVID, un nombre considérable de termes ont été relevés, ce qui enrichit la terminologie actuelle. La méthodologie a été appliquée au français, mais elle peut être reproduite dans d'autres langues. On peut se demander si les prédications qui n'ont pas été étiquetées comme contextes définitoires sont d'une certaine manière des contextes définitoires, car elles fournissent des informations importantes qui permettent de définir les termes. Dans certains cas, il est nécessaire d'élargir la fenêtre de la prédication, car le terme peut avoir été énoncé dans des lignes précédentes. L'analyse des prédications qui ne sont pas des contextes définitoires devient également un élément primordial pour la détection automatique de CD, étant donnée qu'elles constituent un filtre, puisque parmi ces prédications nous pouvons trouver des questions (Quel est le... ?).

Remerciements

Cet article a été partiellment financé par le programme de mobilité de l'Université d'Avignon et le Laboratoire Informatique d'Avignon (France), et par PASPA et PAPIIT IT100822 (Mexique).

Références

Acosta O. et Aguilar C. (2021). Designing a Concept-Mining Model for the Extraction of Medical Information in Spanish. In *Encyclopedia of Information Science and Technology, Fifth Edition*, IGI Global, 856-872

Alarcón R., Bach C. et Sierra G. (2007). Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica. *Revista española de lingüística*, 37, 215-246.

Alarcón R. (2009). Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios. Universitat Pompeu Fabra.

Auger A. (1997). Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles. Université de Neuchâtel.

Bach C. (2005). Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado. *Debate Terminológico*, 1.

Cabré M.T. (1999). La terminología: representación y comunicación. IULA-UPF, Documenta Universitaria, Barcelona.

Espinosa-Anke L. et Saggion H. (2014). Descripción y Evaluación de un Sistema de Extracción de Definiciones para el Catalán. *Procesamiento de Lenguaje Natural*, 53, 69-76.

Estopà R. (2001). Elementos lingüísticos de las unidades terminológicas para su extracción automática", en La terminología científico técnica, editado por Cabré T, Feliu J., IULA-UPF, Barcelona.

Gallego-Hernández D. (2022). Extracción de fraseología especializada basada en corpus: Evaluación de sistemas. *Revista Española de Lingüística Aplicada*.

Malaisé V. (2005). Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels. Université Paris 7 – Denis Diderot.

Meyer I. (2001). Extracting a knowledgerich contexts for terminography: A conceptual and methodological framework. En *Recent Advances in Computational Terminology,* John Benjamins.

Mirapeix M.R., Anke L.E. et Camacho-Collados J. (2020). Definition Extraction Feature Analysis: From Canonical to Naturally-Occurring Definitions. In Workshop on the Cognitive Aspects of the Lexicon.

Navigli R., Velardi P. et Ruiz-Martínez J. (2010). An Annotated Dataset for Extracting Definitions and Hypernyms from the Web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Rebeyrolle J. et Tanguy L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. Cahiers de Grammaire 25. Université de Toulouse-le-Mirail, pp.153-174.

Saggion H. et Gaizauskas R. (2004). Mining on-line sources for definition knowledge. 17th FLAIRS.

Sierra G. (2009). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática* 2, 13-37.

Sierra G., Alarcón R., Aguilar C., Barrón A., Benítez V. et Baca I. (2006). Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología. IX Enc. Int. de Lingüística en el Noroeste.

Storrer A. et Wellinghoff S. (2006). Automated detection and annotation of term

definitions in German text corpora. Proceedings of the Fifth International Conference on Language Resources and Evaluation.
Westerhout E. et Monachesi P. (2007). Extraction of Dutch definitory contexts for eLearning purposes. Proceedings of the Computational Linguistics in the Netherlands (CLIN 2007), 219–34.

Marked Hawkes processes for Twitter data

Andrea Simonetti, Nicoletta D'Angelo, Giada Adelfio

University of Palermo - andrea.simonetti@unipa.it,
nicoletta.dangelo@unipa.it, giada.adelfio@unipa.it.

Abstract
In this paper, we propose to model retweet event sequences using a marked Hawkes process, which is a self-exciting point process where the occurrence of previous events in time increases the probability of further events. The aim is to analyse Twitter data combining temporal point processes theory and textual analysis. Since each retweet event carries a set of properties, we mark the process by different characteristics drawn from the textual analysis, finding that the tone of the description of the Twitter user is a good predictor of the number of retweets in a single cascade.

Keywords: Twitter data, self-exciting, point processes, textual analysis, Hawkes models

Riassunto
In questo articolo, proponiamo di modellare sequenze di eventi retweet utilizzando un processo di Hawkes marcato, che è un processo puntuale self-exciting, in cui il verificarsi di eventi precedenti nel tempo aumenta la probabilità di ulteriori eventi. L'obiettivo consiste nell'analisi di dati Twitter combinando la teoria dei processi temporali e l'analisi testuale. Poiché ogni evento porta con sé un insieme di proprietà, marchiamo il processo con diverse caratteristiche tratte dall'analisi testuale trovando che il tono della descrizione dell'utente Twitter è un buon predittore del numero di retweet in una singola cascata.

Parole chiave: dati Twitter, self-exciting, processi di punto, analisi testuale, modelli di Hawkes

1. Introduction

Point processes are collections of random points falling in some space, such as time and location. Point processes provide the statistical language to describe the timing and properties of events. In geophysics, an event can be an earthquake that is indicative of the likelihood of another earthquake in the vicinity in the immediate future. In the analysis of online social media, events can be user actions over time, each of which has a set of properties such as user influence, the topic of interest, and connectivity of the surrounding network.
In this paper, we propose to model retweet event sequences using a marked Hawkes process, which is a self-exciting point process where the occurrence of previous events in time increases the probability of further events. Indeed,

analysing Twitter data by temporal point processes, each tweet represents an event, which occurs in continuous time. Each event carries a set of properties such as the user influence, the topic of interest, and the connectivity of the surrounding network. Describing Twitter data as a Hawkes model allows us to explain the nature of the underlying process, as well as predict the likelihood and volume of future events.

We follow Rizoiu et al. (2017) for the initial model formulation and fitting, which provides some practical examples drawn from social media data, showing how to model retweet cascades using a Hawkes self-exciting process. They present a design of the memory kernel and results on estimating parameters and predicting popularity, as well as the code and sample event data. Then, we extend their model to textual analysis, replacing the mark of the process, given initially by the number of followers per user, by some of their characteristics elicited from texts. In particular, we perform sentiment analysis on the Twitter description of the user retweeting the original tweet and use a quantitative score based on the resulting emotion as a mark useful for predicting the number of future events, i.e. retweets.

All the analyses are performed through the software Python, for handling the textual data, and the software R Core Team (2022) language for producing the graphical outputs and for fitting the Hawkes model.

The structure of the paper is as follows. Section 2 presents the methodology. In Section 3 we show the application to a dataset concerning retweets of Italian politicians and political parties during November 2017. Finally, conclusions are drawn, and future works are outlined in Section 4.

2. Methods: A marked Hawkes process for information diffusion

A temporal point process is a random process whose realisations consist of the event times T_i, i=1,...,n, falling in R^+. T_i can usually be interpreted as the time of occurrence of the i-th event, and T_i is often referred to as event time.

The simplest class of point processes is the homogeneous Poisson process, where events arrive randomly with the constant intensity λ. However, at least a non-constant intensity is necessary to describe complex processes, such as in an inhomogeneous Poisson process, where the rate of event arrivals is a function of time, i.e., $\lambda = \lambda(t)$. In both models, the events arrive independently. Furthermore, for some applications, the arrival of an event may increase the likelihood of observing events in the near future (e.g. earthquake aftershocks in seismicity, or user interactions in preferential attachment in social networks). For our application, we refer to a class of processes in which the event arrival rate explicitly depends on past events – i.e. *self-exciting processes* – and, in particular, to the *Hawkes process* (Hawkes, 1971).

In general, point processes are characterised by the *conditional intensity function*,

that for the *Hawkes process* has the following form:

$$\lambda(t) = \lambda_0(t) + \sum_{T_i<t} \varphi\,(t - T_i), \quad (1)$$

where $\lambda_0(t)$ is the arrival rate of immigrant events (original tweets) and $\varphi\,(t - T_i)$ is a memory kernel, which describes the self-exciting behaviour of the process (retweets).

In particular, we use the power-law kernel, which appears appropriate for modelling information diffusion in social media (Rizoiu et al.,2017), such as $\varphi(\tau) = \kappa m^{\beta}(\tau + c)^{-(1+\theta)}$.

Furthermore, modelling the retweet cascade as a marked point process means to mark each retweet by the user influence (m), measured for instance as the number of followers of the user emitting the retweet. The power-law kernel depends on four parameters: the virality of the content (κ), the productivity or influence of the user (β), the temporal shift (c), and the temporal decay (θ), all estimated by maximising the log-likelihood over the observed data.

In this paper, we aim at comparing the cascades generated by different tweets on the same topic, in order to assess differences in the popularity prediction, and estimate virality and waiting times, i.e. the diffusion speed. Furthermore, we can predict the total size of each of the diffusion cascades, that is the estimated number of retweets N_∞, as suggested by Mishra et al. (2016).

3. Analysis and results

Italian political retweets data, during November 2017, are downloaded from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QK4MJN.

The dataset consists of a list of more than 100.000 retweets IDs of tweets. For each retweet, we retrieve the user's information and the original tweet through the Twitter API of the Python module *tweepy.* Then, we find 5422 original tweets, posted by 28 Twitter profiles from the top Italian politicians and political accounts. Overall, only 336 cascades contain at least 100 tweets. We select the 10 most popular tweets in the dataset, i.e., the ones with more retweets, by some Italian politicians: Matteo Renzi, Pietro Grasso (2 tweets), Luigi Di Maio (4 tweets), Silvio Berlusconi, Matteo Salvini, and Emma Bonino. The most popular tweet in the dataset is written by Matteo Renzi, with 1821 retweets.

First, we fit the model presented in the previous section, where the mark is given

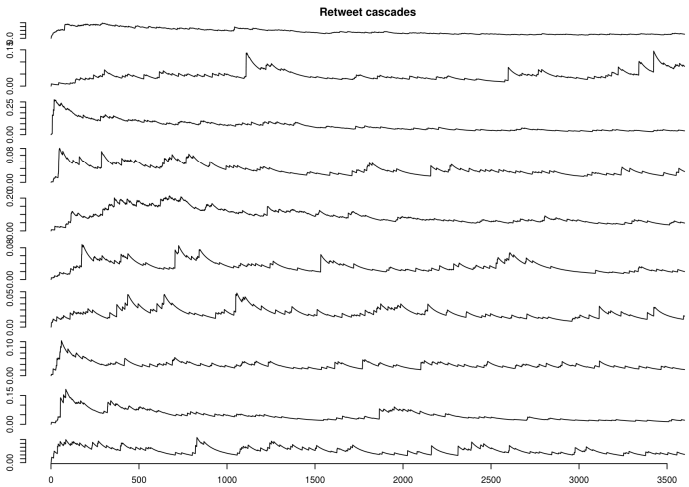
by the number of followers per user. The results of the fitting procedure come in Table 1, and are obtained considering the whole sequence of retweets.

*Table 1. Estimated parameters of the first proposed Hawkes model. The real number of retweets per cascade (*n*), the estimated number of retweets (*N*), and their distance, in percentage.*

Original Tweeter	κ	β	c	θ	n	N_∞	$100*(N_\infty - n)/n$
Matteo Renzi	0.25	0.43	89.86	0.48	1821	1840	1.04
Pietro Grasso	1.00	0.33	269.25	0.49	1303	1335	2.46
Luigi Di Maio	0.84	0.59	573.81	0.68	977	981	0.41
Luigi Di Maio	0.80	0.24	196.55	0.40	940	995	5.85
Silvio Berlusconi	0.32	0.45	149.85	0.50	935	963	3.00
Luigi Di Maio	1.00	0.26	176.75	0.46	838	859	2.51
Matteo Salvini	1.00	0.30	282.08	0.46	803	831	3.49
Pietro Grasso	1.00	0.42	392.52	0.57	786	799	1.65
Luigi Di Maio	0.67	0.33	278.43	0.45	737	777	5.43
Emma Bonino	1.00	0.36	316.61	0.53	712	730	2.53

In Figure 1, we plot part of the observed event series and the corresponding intensity function with fitted parameters. We show only the first hour, even though the model is fitted on the whole sequences' length.

Figure 1. Estimated intensity functions over one hour for the 10 retweet cascades. The parameters used are those in Table 1.



Looking at Table 1, we can state that, overall, the estimated parameters appear similar over the different analysed cascades, and the predicted number of retweets almost always resembles the real value, with very small errors, reaching a maximum of 5.85%.
Then, we move to include textual information in the model. In other words, we account for some characteristics of the user, instead of the number of followers. In detail, we compute the anger, joy, fear, and sadness scores of the Twitter user's description, by means of a neural network model, called UmBERTo, that was pre-trained on a novel benchmark corpus of Italian Twitter posts, called FEEL-IT, annotated with the four basic emotion (https://huggingface.co/MilaNLProc/feel-it-italian-emotion). The model is available through the open-source Python library (https://github.com/MilaNLProc/feel-it).
The resulting score variables range between 0 and 1, representing the probability of the tone of the description being one among the 4 considered emotions. We also assign the category "neutral" to all the descriptions of less than 2 words. The underlying assumption is that the textual characteristics of the description of the user, and therefore the overall mood of the user, influences the number of retweets. We, therefore, fit two additional Hawkes models, with the anger score and the joy score in place of the number of followers, and compute the χ^2 statistics on the number of predicted retweets in the 10 cascades. The results come in Table 2.

Table 2. χ^2 statistics on the number of predicted retweets in the 10 cascades, for the three fitted models.

Mark	χ^2 statistics
Number of followers	9.00525
Anger score	8.45834
Joy score	9.97102

From Table 2, we can first notice that the considered quantitative variables resulting from the textual analysis of the users' descriptions only achieve similar error terms, if compared with those obtained marking the temporal process by the number of followers. Furthermore, we notice that the anger score actually predicts better the number of retweets in the single cascades. Finally, Figure 2 displays the distributions of the estimated parameters for the three models fitted to the Twitter data.

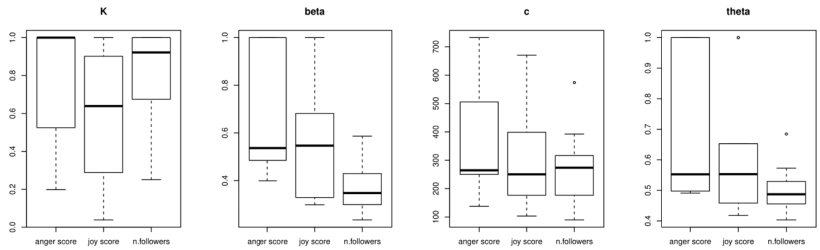


Figure 2. Distributions of the estimated parameters for the three models fitted to the Twitter data.

It is evident that the choice of the mark has an impact on the estimates. The virality of the content, represented by the κ parameter, of the anger score is similar to the one of the numbers of followers, while the joy score seems to lead to a lower virality. This is a reasonable result, as it enforces the idea that the angrier the user, the more visible the tweet and the consequent retweets. Then, the mark parameter β indicates the productivity or influence of the user. The estimates for the tones of the description of the user (anger and joy) are larger if compared to that of the number of followers, indicating a higher effect on the number of subsequent events in the temporal process. The higher distribution of the β parameter is the "anger score" one. Finally, the temporal shifts and decays (parameters c and θ) also show some degree of variability among the models: the waiting time is larger when considering only the number of followers as the mark in the model, and this is also reflected in the temporal decay, which is lower the ones for models marked by the anger and joy scores.

4. Conclusions

In this paper, we have proposed a marked Hawkes process to model retweet event sequences where the occurrence of previous events in time increases the probability of further events. This has allowed us to explain the nature of the underlying process. The most relevant result is combining textual analysis results with the context of temporal point process theory, and drawing relevant conclusions. In particular, we have found that the tone of the description of the Twitter user, is a good predictor of the number of retweets in the single cascade, comparable to the number of followers. These represent only preliminary results, and in future, we aim at analysing larger samples. Furthermore, considering multiple cascades in future would allow us to extend the proposed model, by introducing a nonzero occurrence rate of original tweets, also describing background events with independent occurrences of the previous events.

Moreover, as each event carries a set of properties such as the user influence, the topic of interest, and the connectivity of the surrounding network, in future works, we could as well introduce other covariates. In the context of the Twitter data, external covariates may consist of several textual features, such as similarities among texts, analysis of word co-occurrences, etc. From a methodological point of view, the Hawkes model could be specified in order to include multiple covariates, in a generalised linear model fashion.

Finally, while in this paper we have only focused on the characteristics of the users, the textual analysis can also provide some information about the main characteristics of the tweets, such as the level of information provided, the writing formalism, the loudness used, and so on. In the future, we aim at using this information and projecting it in a conventional space, by multivariate analysis approach, such as the Multidimensional Scaling (MDS), where each point still represents the tweet/event, now identified both by the temporal and spatial locations. This would allow us to further extend the previous methodology to the spatio-temporal context, for assessing the distribution and interaction of events and characterising their 'distance' in the new space.

References

Bianchi, F., Nozza, D., and Hovy, D. (2021). Feel-it: Emotion and sentiment classification for the italian language. In Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis

Giglietto, F.; Rossi, L. (2017), "Retweets of Italian politicians and political parties during November 2017", https://doi.org/10.7910/DVN/QK4MJN, Harvard Dataverse, V2, UNF:6:OED72hOAvWk3AH1Ga3otqw== [fileUNF]

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83-90.

Mishra, S., Rizoiu, M. A., and Xie, L. (2016). Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 1069-1078).

R Core Team (2022). A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Rizoiu, M. A., Lee, Y., Mishra, S., and Xie, L. (2017). A tutorial on Hawkes processes for events in social media. *arXiv preprint arXiv:1708.06401*.

Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison: à propos de l'étude stylistique des lettres de Pierre Damien

Valérie Thon[1], Laurent Vanni[2], Dominique Longrée[3]

[1]Université de Paris – valerie.thon@u-paris.fr
[2]Université Côte d'Azur – laurent.vanni@unice.fr
[3]Université de Liège – dominique.longree@uliege.be

Abstract
To carry out a complete and reliable morphosyntactic labeling of Latin texts is a particularly time-consuming task. It is therefore necessary to choose wisely the texts to be included in a labelled comparison corpus when one wishes to study the intertextual distances between a given author, in particular a medieval one, and his predecessors. A stylistic research on the letters of Peter Damian (11th century) was the occasion to question the methods to be implemented to operate this selection. The intertextual distances were first computed on the forms using additive tree analysis. The results were then compared to the predictions of the deep learning, attributing with variable recognition rates passages of Damian to various authors of the comparison corpus. Where ADT relies primarily on the lexicon, the Convolutional Neural Network takes into account morphosyntactic parameters, with strong areas of activation suggesting a recognition of linguistic patterns that Damian shares with some of his predecessors.

Keywords: ADT, deep learning, prediction, lemmatization, morphosyntax, Peter Damian

Résumé
Réaliser un étiquetage morphosyntaxique complet et fiable de textes latins est une tâche particulièrement chronophage. Il s'agit dès lors de choisir à bon escient les textes à intégrer à un corpus de comparaison étiqueté lorsque l'on désire étudier les distances intertextuelles entre un auteur donné, en particulier un auteur médiéval, et ses devanciers. Une recherche stylistique sur les lettres de Pierre Damien (XI^e siècle) a été l'occasion de s'interroger sur les méthodes à mettre en œuvre pour opérer cette sélection : les distances intertextuelles ont été d'abord calculée sur les formes à l'aide d'analyses arborées ; les résultats ont été ensuite comparés aux prédictions du deep learning, attribuant, avec des taux de reconnaissance variables, des passages de Pierre Damien à divers auteurs du cor-

pus de comparaison : là où l'ADT semble s'appuyer essentiellement sur le lexique, le Convolutional Neural Network prend mieux en compte des paramètres morphosyntaxiques, les zones d'activation fortes suggérant une reconnaissance de motifs linguistiques que Damien partagerait avec certains de ses prédécesseurs.

Mots clés: ADT, deep learning, prédiction, lemmatisation, morphosyntaxe, Pierre Damien

1. Introduction

Le deep learning offre-t-il des instruments pour compléter les méthodes de l'ADT lorsqu'il s'agit d'établir un corpus de comparaison en vue d'une analyse stylistique ? Nous tenterons de répondre à cette question dans le cadre d'une étude textométrique plus large de la langue et du style de Pierre Damien (1007-1072/73), ermite 'réformateur' du Moyen Âge central, auteur de 180 lettres latines. Dans le cadre du projet doctoral de V. Thon, 12 d'entre elles (33.396 occurrences), couvrant la totalité de sa carrière ecclésiastique, ont été traitées à cette fin selon les méthodes du LASLA (lemmatisation et analyse morphosyntaxique semi-automatique, avec une vérification manuelle complète). Complétant les méthodes de la stylistique traditionnelle, ce traitement permet non seulement une analyse textométrique fine de l'évolution de la langue et du style de l'auteur à travers sa carrière, mais aussi de le confronter à ses devanciers pour identifier ceux qui auraient influencé sa technique d'écriture. La mesure de cette 'influence diachronique' implique d'aller au-delà du domaine lexical, trop sensible à la thématique des divers textes, et d'examiner les caractéristiques morphosyntaxiques qui présentent une plus grande indépendance vis-à-vis de celle-ci. Pour ce faire, il s'agit de disposer d'un corpus de comparaison lemmatisé et étiqueté morphosyntaxiquement. Pour la période classique, la banque de textes du LASLA fournit un corpus étiqueté important, mais qui ne couvre pas l'ensemble des textes susceptibles d'avoir influencé Pierre Damien. Il est également nécessaire de pouvoir confronter ses lettres aux écrivains tardo-antiques ou médiévaux, tels les Pères de l'Église ou les auteurs carolingiens. Or ce corpus « tardif » est de taille trop considérable pour être traité intégralement. Le défi principal est dès lors d'identifier –sans étiquetage préalable– les auteurs tardo-antiques et médiévaux avec lesquels Pierre Damien présente des affinités et qui seraient ainsi des candidats par excellence à un étiquetage fin. Pour ce faire, nous proposons de croiser ADT et deep learning (module de deep learning – CNN fondé sur Hyperbase Web) sur un grand corpus de textes non-traités (40 auteurs au total).

2. Les méthodes du LASLA: fiables, mais chronophages

Le traitement des textes latins avec les méthodes du LASLA consiste à procéder à une lemmatisation et à une analyse morphosyntaxique complète de chacune des formes rencontrées, en utilisant une procédure semi-automatique : les analy-

ses que propose le logiciel font l'objet d'une sélection et d'une vérification systématique par un philologue confirmé qui peut compléter ou corriger les informations fournies. Ce travail a le mérite d'assurer une très grande fiabilité aux données contenues dans les fichiers du LASLA, mais, même si cette opération se fait en ligne grâce à l'interface LEI (LASLA Encoding Initiative), celle-ci requiert un temps de travail important (de l'ordre de 100 à 150 heures pour 20.000 mots). Le travail est d'autant plus chronophage quand le texte comprend des termes tardo-antiques ou médiévaux inconnus du dictionnaire classique. En collaboration avec le projet Collatinus, le LASLA a pu trouver une alternative à la procédure LEI : en s'appuyant sur un dictionnaire enrichi, Collatinus-LASLA permet de traiter un peu plus rapidement les textes, mais sans pour autant éliminer une intervention manuelle importante (Verkerk *et alii*, 2020). Le LASLA s'est par ailleurs tourné à diverses reprises vers l'utilisation de lemmatiseurs ou étiqueteurs complétement automatiques, mais malheureusement sans arriver à une fiabilité suffisante pour une étude stylométrique fine (Longree et Poudat, 2010). Dans ce contexte, il s'agit donc de choisir avec circonspection les textes à retenir pour constituer le ou les corpus de comparaison, textes devant faire l'objet d'un traitement avec la procédure LEI. Cette sélection peut heureusement s'appuyer sur des analyses préalables portant sur des textes non lemmatisés. En effet, le latin étant une langue flexionnelle, les formes contiennent dans leurs terminaisons une information morphosyntaxique importante que les méthodes de l'ADT peuvent prendre en compte, mais seulement jusqu'à un certain point. Le recours au Deep Learning permet, comme on va le voir, d'aller un peu plus loin.

3. Pour une confrontation méthodologique : ADT et deep learning

3.1. Méthodologie : la constitution du corpus de comparaison

Pour étudier le style de Pierre Damien, nous avons constitué un corpus de 40 auteurs susceptibles de l'avoir influencé. En nous appuyant sur les informations fournies dans l'édition critique de Reindel (1983-1993) et sur le catalogue (fin XI^e^ – début XII^e^ siècle) de la bibliothèque de Fonte Avellana, monastère où Pierre Damien fut prieur à partir de 1043 (Martini, 2002), nous avons veillé à ce que le corpus comprenne à la fois des auteurs que Pierre Damien a certainement connus (voire lus), et des auteurs avec lesquels aucun lien particulier n'est supposé a priori. Ceux-ci sont répartis de façon équitable sur trois périodes, correspondant à 3 modèles *Hyperdeep* entraînés (Tableau 1) : Antiquité (« Période 1 »), Antiquité tardive et haut Moyen Âge (« Période 2 ») et Moyen Âge central jusqu'à l'époque de Pierre Damien (« Période 3 »).

Tableau 1. Caractéristiques du corpus de comparaison divisé en 3 périodes

	Nombre d'occurrences	Nombre d'auteurs	Précision du modèle *Hyperdeep*
Période 1 (IIIe av. – Ie apr.)	1.001.582	12	98.05%
Période 2 (IIe apr. – VIIIe apr.)	1.820.007	15	95.78%
Période 3 (IXe apr. – XIe apr.)	951.511	13	95.56%

Cette sélection obéit à un certain nombre de contraintes : au sein d'une même période, les textes retenus présentent des tailles approximativement identiques, pour éviter que des tailles trop différentes puissent biaiser les résultats ; seuls des textes comprenant plus de 30.000 mots ont été retenus ; chaque modèle entraîné ne dépasse pas la limite de 20 classes. Pour construire un corpus riche et varié, les œuvres choisies se distinguent les unes des autres non seulement d'un point de vue chronologique, mais également par le genre et le style : les textes de comparaison sont donc théologiques, ecclésiastiques, épistolaires, rhétoriques, 'romanesques', historiographiques, philosophiques.... Nous avons toutefois écarté pour le moment tout texte poétique du corpus, parce que la structuration linguistique plus libre des textes poétiques par rapports aux textes prosaïques risquait d'altérer les résultats du deep learning. Enfin, nous avons harmonisé les graphies, ponctuations et mises en pages (les ponctuations faibles ont été supprimées et toutes les ponctuations fortes homogénéisées sous forme de points).

3.2. Les résultats de l'ADT

Dans un premier temps, nous avons soumis les trois corpus de comparaison à un calcul de distances intertextuelles. Dans le cadre imparti ici, nous ne pourrons présenter que les résultats de la « Période 2 ». Les distances ont été successivement calculées selon la méthode Brunet (fréquence des formes) puis la méthode Jaccard (présence/absence des formes) telles qu'implémentées dans Hyperbase Web Edition (avec représentation arborée d'après le procédé de regroupement Neighbour Joining appliquée à une représentation de type radial).

La première méthode, celle de Brunet (Figure 1), rapproche Pierre Damien de Jérôme (IXe – Ve siècle) et de Grégoire le Grand (VIe – VIIe siècle). C'est avant tout le genre pastoral qui prédomine chez eux : Jérôme et Grégoire ont tous les deux rédigé un commentaire sur les prophéties bibliques d'Ézéchiel, textes inclus dans le corpus, et Grégoire est également l'auteur d'une *Règle Pastorale*, destinée à montrer aux évêques de son temps comment ils doivent guider les fidèles dont ils ont la charge. La proximité de Pierre Damien avec ceux-ci n'est pas surprenante. Celui-ci fait figure de maître spirituel à l'esprit pastoral, veillant à ce que ses moines s'orientent toujours vers Dieu, et n'hésitant pas à adresser divers con-

seils à ses contemporains, hommes et femmes, clercs et laïcs.
La méthode « Jaccard » (Figure 2) situe les lettres de Pierre Damien sur une branche qui regroupe Tertullien (IIe – IIIe siècle), Hilaire (IVe siècle) et Alcuin (VIIIe – IXe siècle).

Figure 1. Calcul des distances Brunet

Figure 2. Calcul des distances Jaccard

Les textes de ces trois auteurs s'organisent tous autour de l'apologie : Tertullien est représenté par ses *Prescriptions contre les hérétiques*, son *Apologétique* (sa grande défense de la religion chrétienne) et son *De l'idolâtrie* ; Hilaire est présent dans le corpus par son *De Trinitate*, défense du dogme chrétien de la trinité ; Alcuin est essentiellement représenté par un traité voué au même sujet (*De fide trinitatis*). Le fait que Pierre Damien se range parmi ces auteurs pourrait de prime abord ne pas surprendre : défenseur de l'idéologie « grégorienne », il lance régulièrement des attaques contre toute pratique à son avis inacceptable. Mais ce déplacement thématique d'une méthode à l'autre pose néanmoins question. Or les textes rapprochés par la méthode Jaccard ont un autre point commun : les textes pris en compte pour Tertullien, Alcuin et Pierre Damien présentent une taille plus réduite que les autres textes du corpus de la période 2. Les résultats obtenus peuvent donc aussi s'expliquer par la grande sensibilité de la méthode Jaccard à la taille des textes : ce ne sont donc pas nécessairement les apologètes qui se rapprochent de Damien en raison d'un vocabulaire partagé, mais plutôt les autres textes qui repoussent tant les apologètes que Damien de par des absences lexicales liées à la faible taille de leurs partitions respectives. Il n'en reste pas moins que, dans les deux méthodes les spécificités lexicales jouent un rôle déterminant, alors que le rôle de la morphosyntaxe est quant à lui totalement occulté. Or des travaux antérieurs ont bien mis en évidence (Vanni et al., 2018a et 2018b) la sensibilité du deep learning à la morphosyntaxe. Nous avons donc vérifié s'il en était de même ici.

3.3. Les résultats du deep learning

Les 12 lettres de Pierre Damien faisant partie de la thèse de V. Thon, –supposées former un échantillon représentatif–, ont d'abord été proposées une par une au modèle entraîné sur le corpus « Période 2 », en lui donnant comme tâche d'attribuer ces lettres à un des auteurs du corpus. Afin de vérifier la stabilité des résultats sur une population plus large, nous avons ensuite procédé de même avec la quasi-totalité de ses lettres en un seul fichier (édition Migne ; 158 lettres étaient connues à l'époque de cette édition). Les 12 lettres ont été souvent attribuées en premier lieu à Bède le Vénérable (VII^{e} – VIII^{e} siècle), avec des taux de reconnaissance particulièrement élevés, puis à Grégoire le Grand, proche de Damien avec la méthode Brunet. Ambroise (IV^{e} siècle) et Augustin (IV^{e} – V^{e} siècle) sont également susceptibles d'avoir un taux important, mais sont globalement moins présents. Le résultat pour l'ensemble des 158 lettres semble confirmer ce constat : Grégoire le Grand (23%) y occupe la première position, suivi de près par Bède (22%) et Ambroise (13%) ; les autres auteurs du corpus de comparaison ne s'y élèvent pas au-dessus de 10%.

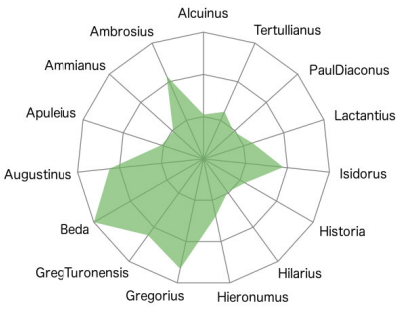


Figure 3. Attribution de la Lettre 114

L'attribution des textes de Pierre Damien à Grégoire, Ambroise ou Augustin n'est pas nécessairement étonnante, étant donné que ces trois derniers sont tous reconnus comme de grands Pères de l'Église dont les écrits ont exercé une influence indéniable sur tout le Moyen Âge latin. Le lien avec Bède, en revanche, est plus surprenant, d'autant qu'il n'est pas suggéré par les arborées de l'ADT. Il est représenté dans le corpus par deux textes plutôt historiques : *Histoire ecclésiastique du peuple anglais* et *De la raison du temps*. La reconnaissance semble s'appuyer non seulement sur l'idée d'un vocabulaire partagé (le champ sémantique du religieux et du temporel, par exemple, est présent chez tous les deux), mais également sur des structures d'ordre morphosyntaxique. Remarquons en effet que le deep learning semble avoir une sensibilité à la morphosyntaxe. Parmi les motifs linguistiques repérés pour Bède, par exemple, nous trouvons régulière-

ment des expressions du type « *qui nimirum, qui dudum, qua uidelicet, quae nimirum, quae statim, cui uidelicet,…* ». La figure 4 fournit un exemple des taux d'activation important mis en évidence pour la séquence *cui uidelicet ad corrigendum* grâce au calcul du TDS (Vanni et al. 2018b).

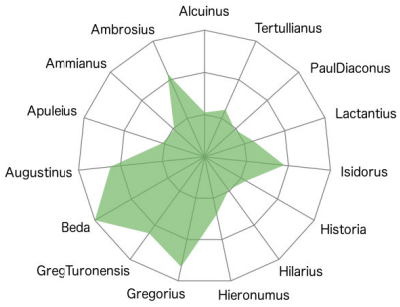


Figure 4. Exemple de motifs repéré par le deep learning de la Lettre 114

Une recherche sur la distribution des deux expressions *qui uidelicet* et *cui uidelicet* dans le corpus « Période 2 » montre que celles-ci sont fortement spécifiques de Bède. Le deep learning aurait-il reconnu plus largement au sein des lettres de Pierre Damien une structure syntaxique « pronom relatif + adverbe », qui serait également caractéristique du style de Bède ? Lemmatiser un ou plusieurs textes de Bède permettrait de vérifier cette hypothèse par les méthodes éprouvées de l'ADT.

4. Évaluation

Cette recherche visait à croiser ADT et méthode prédictive du réseau de neurones dans le cadre d'une étude textométrique du corpus épistolaire de Pierre Damien. L'objectif était de détecter d'éventuels modèles stylistiques parmi les textes de ses devanciers, en vue d'identifier des candidats à une lemmatisation semi-automatique fine. D'une part, l'ADT a perçu un lien étroit entre Pierre Damien et le genre pastoral (Grégoire le Grand et Jérôme), mais suggérait également un rapport avec l'apologie (Tertullien, Hilaire, Alcuin), alors que celui-ci est essentiellement dû à la trop grande sensibilité de la méthode Jaccard à la taille des textes et donc aux absences de formes. De son côté, le deep learning, ne prend en compte lui que des présences et surmonte cette difficulté. Ainsi, il a identifié une proximité non seulement avec Grégoire le Grand, mais aussi avec Bède le Vénérable. Ce résultat était inattendu : les zones d'activations suggèrent que l'attribution s'appuie non seulement sur des critères lexicaux, mais aussi morphosyntaxiques, le CNN semblant capable d'identifier des catégories grammaticales comme celles du relatif ou de l'adverbe, en ne s'appuyant que sur les formes seules. Cette recherche montre une nouvelle fois tout l'intérêt d'un dialogue entre

les méthodes, dialogue qui pourra se prolonger une fois les textes de Grégoire et de Bède ayant été lemmatisés et étiquetés pour permettre une analyse textométrique fine.

Références

Brunet É. et Vanni, L. (2019). *Deep learning* et authentification des textes. *Texto ! Textes et cultures*, vol. 24.1 : 1-34.

Longrée D. et Poudat, C. (2010). New Ways of Lemmatizing and Tagging Classical and post-Classical Latin: the LATLEM project of the LASLA. In P., Anreiter et M., Kienpointner (Eds.), *Proceedings of the 15th International Colloquium on Latin Linguistics. Innsbruck.*

Martini P. S. (2002). L'inventario del secolo XII della biblioteca di Santa Croce di Fonte Avellana. In Gatto L. et Martini P. S. éditeurs, *Studi sulle società e le culture del Medioevo per Girolamo Arnaldi*, All'Insegna del Giglio.

Reindel K. (éditeur, 1983-1993). *Die Briefe des Petrus Damiani. Teil 1 – 4*. Monumenta Germaniae Historica, Die Briefe der deutschen Kaiserzeit.

Vanni L, Mayaffre D. et Longrée, D. (2018a). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. *JADT 2018* : 459-466.

Vanni L., Ducoffre M., Mayaffre D., Precioso F., Longrée D., Elango V., Santos N., Gonzalez J., Galdo L. and Aguilar C. (2018b). Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) : 548-557.

Verkerk P., Ouvrard P., Fantoli M. et Longrée D. (2020). L.A.S.L.A. and Collatinus: a convergence in lexica. In L., Tesconi (Ed.), *Studi e saggi linguistici* 2020(1) (pp. 1-26). Edizioni ETS.

Comparison of two annotation schemes to derive offensiveness scores in HurtLex

Alice Tontodimamma[1], Stefano Anzani[1], Valerio Basile[2], Marco Stranisci[2], Lara Fontanella[1]

[1]University of Chieti-Pescara – alice.tontodimamma@unich.it
s.anzani92@gmail.com lara.fontanella@unich.it
[2]University of Turin – valerio.basile@unito.it
marcoantonio.stranisci@unito.it

Abstract
The recent proliferation of social media has been accompanied by a similar widespread increase in the frequency of hostile online messages and inflammatory speeches. Social media are nowadays seen as a suitable arena for ill-tempered debates to be conducted, with the frequent use of insulting terms and other offensive language. Lexical resources containing specific negative words have been widely employed to detect uncivil communication. This work describes the development, implementation and comparison of two annotation schemes to derive offensiveness scores in a lexicon of hate words, namely HurtLex. Since determining a given expression's offensiveness is a highly subjective matter, we decided to use two different methods to derive offensiveness scores. The first method consists in a rating scale annotation such that each word has to be assessed on a 5-point Likert scale and to derive a final score we explicitly model the item response probability trough a unidimensional Graded Response Model (GRM). The second one is the Best-Worst scaling method. We analyse whether these two different methods lead to correlated scores. We propose a qualitative-quantitative analysis of the terms that have opposite scores. The two scoring systems are evaluated on two annotated corpora. The introduction of an offensiveness level score in HurtLex could be helpful as part of a lexicon-based approach to track offensive and hateful content.

Keywords: hate speech detection, lexicon-based approach, Item Response Theory Model, Best-worst scaling, comparison annotation schemes.

1. Introduction

Any form of communication that disparages a person, or a group of people, on the basis of their race, their colour, their ethnicity, their gender, their sexual ori-

entation, their nationality, or their religion is regarded as hate speech. The recent huge expansion in the use of social media has been accompanied by a proportionate increase in the online diffusion of hate speech and hate-based propaganda, to the point where this is now recognised to be a serious problem: consequently, action is now under way at international level to address the problem and to develop effective means of limiting and counteracting the spread of this unpleasant phenomenon. Attempts to study and define hateful and offensive communication or to make a linguistic description to identify its characteristics begin from the lexical level. The presence of offensive and threatening language in a text could represent a feature to identify such a text as uncivil. Many authors (see Schmidt and Wiegand, 2017, for a comprehensive review) therefore use lists of offensive expressions (e.g. slurs, insults, abusive and swear words) as lexical resources for this purpose; the majority of these resources, however, have been compiled for the English language, and until recently few such lists existed in Italian. A list of Italian taboo words and expressions, intended for use in mining offensive language on social media, has been proposed by Maisto et al. (2017). Italian offensive and hateful words can also be found in HurtLex (Bassignana et al., 2018), a multilingual lexicon built semi-automatically from the originally handcrafted Italian lexicon `Words as Weapons' (De Mauro, 2016). There is also an updated version of HurtLex, the revised HurtLex dictionary (Tontodimamma, 2021), which concentrates on the Italian entries, revising the terms listed in HurtLex and using an Item Response Theory model to create an offensiveness score for each lexical item.

In this paper we therefore set out to describe the development and implementation of two annotation schemes intended to create a more reliable offensiveness score, both using the revised HurtLex as their lexicon, together with a comparison between the two schemes.

Researchers in different disciplines, including social sciences and computational linguistics, often rely on rating scales, when manually annotating data with quantitative or qualitative information. The annotations for an item from multiple respondents are usually averaged to obtain a real-valued score for that item. Different authors (see Kiritchenko and Mohammad, 2017; Poletto et al., 2019, and references therein) propose different annotation schemes, such as Best-Worst scaling. In this paper, we analyse whether these two different methods lead to correlated scores. We produce offensiveness intensity ratings for 168 Italian terms extracted from the revised Hurtlex dictionary, using both methods, by aggregating responses from several independent annotators. We propose a qualitative-quantitative analysis of terms in disagreement. The two different scoring systems are evaluated on two annotated corpora. The introduction of an offensiveness level score in HurtLex can be helpful as part of a lexicon-based approach to tracking offensive and hateful content.

2. Annotating for Offensiveness

We annotated 168 terms for offensiveness with both rating scale and best-worst scaling. Task 1 was done using Label Studio[1] as a platform. While the second task was done using Lite Scale[2], a tool specifically created for the manual annotation of linguistic data with the Best Worst Scaling method. The annotation process was carried out by thirteen trainees (two male, eleven female, students on the Sociology degree course) who are engaged in an internship program in the Computational Social Research Lab[3]. The average time taken for the rating scale annotation is 30 minutes, while for the second one is 50 minutes.

2.1. Terms

The term list comprises 168 words drawn from the revised HurtLex dictionary. TWITA (Basile et al. 2018) is a large-scale collection of Italian-language tweets that has been made available for research purposes. We have computed the frequency of all terms used in this collection over a period of one month (28 September to 28 October 2021) that also appear in the revised HurtLex dictionary, and selected the 200 most frequently used words. In the case of words that appear in several different inflections, we have kept only the most frequently appearing inflection for evaluation. The final result (with the different inflections removed) is a subset of 168 words.

2.2. Task 1: Annotating with Rating Scale and offensiveness level estimation through the Graded Response model

In this work, we adopted a rating scale annotation such that each word was to be assessed on a 5-point Likert scale. In particular, in this annotation task each rater states how offensive that word is on a scale ranging from 1 to 5. Rating scales present several challenges, such as the difficulty in maintaining inter- and intra-annotator consistency. To overcome this consistency issue, instead of simply taking the average of the individual ratings, we rely on an item factor analytic model suitable for ordinal responses.

For the present study, the final score for each entry word was determined through a unidimensional graded IRT model. The graded response model (GRM), introduced by Samejima (1969), represents a family of appropriate models when dealing with ordered categories. The GRM is a generalization of the two-parameter logistic (2PL) model. Our analysis relates the latent trait to the words while the

[1] https://labelstud.io/.
[2] http://gingerbeard.alwaysdata.net/litescale-web/home.
[3] http://csrlab.unich.it/.

so-called item parameters are related to respondents. In particular, we assume that the score assigned to a chosen term depends on its intrinsic latent level of offensiveness and on some respondent's parameters. The respondent's parameters represent the weight of the rater's evaluation in the composite index of offensiveness and the threshold values between consecutive categories along his/her continuous evaluation scale. Formally, denoting by Xij the score assigned by rater *j* to word *i* and assuming a 2-parameter logistic model (2PL), the probability that the score will fall into category c=0,1,…,C is given by

$$P(X_{ij} = c|\theta_i, \alpha_j, d_j) = P(X_{ij} \geq c|\theta_i, \alpha_j, d_j) - P(X_{ij} \geq c+1|\theta_i, \alpha_j, d_j). \quad (1)$$

where θ_i denotes the level of offensiveness of word i, α_j is the factor loading, or discrimination parameter, for rater *j*, and finally $d_j = (d_{j1}, \dots, d_{j(C-1)})'$ is the vector of ordered threshold parameters $-\infty \leq d_{j1} \leq \cdots \leq d_{j(C-1)} \leq \infty$. In our case, C=5.To estimate the model parameters, we used the mirt library (Chalmers, 2012) in the language and environment for statistical computing R (R Core Team 2018). Estimation was carried out via the Expectation-Maximisation algorithm.

2.3. Task two: Annotating with Best Worst

Best-Worst Scaling (BWP) is a comparative annotation model developed by Louvriere and Woodworth (1991). This method consists of presenting annotators *n* entries at a time, where *n> 1* and generally *n* = 4. For the *n* items, the annotators are asked to state which one is the "best" in terms of the property you are interested in testing, and which one is the "worst". When working on 4-tuples, best–worst annotations are particularly efficient because by answering these two questions, the results for five out of six item–item pair-wise comparisons become known. All items to be rated are organized in a set of m 4-tuples (*m≥ N*, where *N* is the number of items) so that each item is evaluated several times in diverse 4-tuples. Once the *m* 4-tuples are annotated, one can compute real-valued scores for each of the items using a simple counting procedure (Kiritchenko and Mohammad, 2017).

We created a set of tuples made up by four words (4-tuples), grouped so that each word is repeated four times in the dataset, combined with three different words each time. Then we provided raters with a set of 4-tuples: for each 4-tuple they were asked to point out the most offensive and the least offensive of the four. Raters have thus seen each word four times but have had to compare it with different words every time[4].

[4] http://valeriobasile.github.io/Best-worst-scaling-and-the-clock-of-Gauss/.

3. Task Annotation Results

Figure 1 shows the scores obtained using the two methods of scoring. As can be seen from the scatter plot, the scores are closely correlated: it was therefore not necessary to carry out a qualitative-quantitative analysis of the terms with opposite scores.

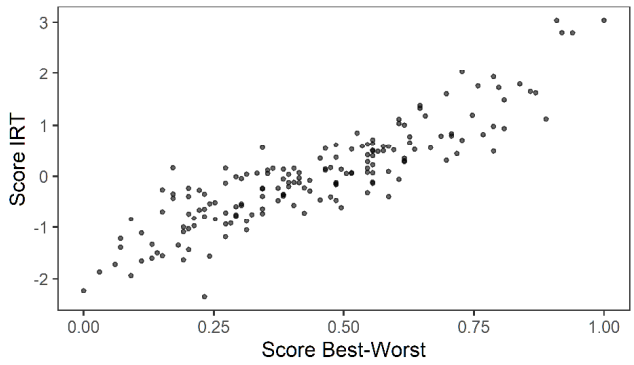


Figure 1. Scatter-plot of the scores obtained through the two scoring systems.

4. Classification tests with different schemes at comparison

Having described the process and results for each task, we will now evaluate these two different scoring systems on abusive language detection tasks on two corpora.

4.1. The analysed corpora

We selected two annotated corpora as benchmarks from the 2020 edition of EVALITA, the recurring evaluation campaign of language technologies for Italian. The corpora were created in the context of two shared tasks, namely Hate Speech Detection (HaSpeeDe) and Automatic Misoginy Identification (AMI), both running for the second time after successful editions at EVALITA 2018. The HaSpeeDe 2020 corpus (Sanguinetti et al., 2020) comprises a training set of 6,839 Italian tweets, including the corpus used for the 2018 edition of the task and extended with data from project "Contro l'Odio" (Capozzi et al., 2019). The tweets are annotated in a binary fashion according to the presence of hate speech (HS, 2,766 tweets) or its absence (not HS, 4.073 tweets). The newly added data also form one of the two test sets of this task, comprising 1,263 tweets (622 HS, 641 not HS). An additional test set is provided, made of 500 Italian news articles (181 HS, 319 not HS) for cross-domain HS detection benchmarking. The AMI 2020 corpus (Fersini et al., 2020) comprises 5,000 Italian tweets as the training set, of which 2,337 are annotated as containing misogynous language, and 1,000 for the test set (500 misogynous, 500 not misogynous).

4.2. Classification Results

We performed classification experiments on the corpora described in the previous section employing a straightforward state-of-the-art model. We encoded each text in the training and test sets with AlBERTo, a neural language model based on BERT (Devlin et al., NAACL 2019) and pre-trained on Italian tweets (Polignano et al., 2019). In particular, given a text preprocessed with the AlBERTo tokenizer (based on WordPiece), we considered the last layer of the output of the encoder as a sequence of 768-dimensional embedded representations of the words in the text. The sequences were capped at 50 elements of length. In order to test the impact of the information provided by the scores introduced in this paper, we further encoded the words in the tests as 17-dimensional vectors, each corresponding to a HurtLex category, similarly to Koufakou et al., 2019. In this kind of encoding, the scores for each word belonging to a category of HurtLex are the values of the corresponding dimension in the embedding. We produced three versions of these embeddings, one for each score (BWS and IRT), and one version where the words are simply counted for each category, as if their scores were always 1, in order to test the impact of the actual score in contrast with the impact of the HurtLex lexicon and categories on the final classification outcome. The HurtLex embeddings are concatenated to the AlBERTo embeddings and fed into a 16-unit Long Short-term Memory (LSTM), followed by a 0.5 drop-out layer and a single-neuron output layer with sigmoid activation. The model is trained for 10^{-3} epochs with Adam optimizer (Kingma and Ba, 2014), learning rate of , and a batch size of 64. The tests were repeated 10 times and the results were averaged to balance the effect of the random initialization of the weights of the neural network.

Table 1. Classification results on HateSpeeDe, Twitter test set, HaspeechDe News Test set, AMI, in terms of positive class precision, recall, and F1-score.

	HaSpeeDe, Twitter test set			HaSpeeDe, News test set			AMI, test set		
Model	**precision**	**recall**	**F1-score**	**precision**	**recall**	**F1-score**	**precision**	**recall**	**F1-score**
baseline	0.687	0.794	0.735	0.767	0.398	0.516	0.714	0.888	0.789
HurtLex count	0.700	0.765	0.725	0.757	0.398	0.500	0.731	0.883	0.799
HurtLex BWS	0.680	0.821	0.742	0.712	0.433	0.530	0.714	0.891	0.791
HurtLex IRT	0.685	0.822	0.745	0.755	0.414	0.525	0.721	0.893	0.796

As it can be observed in Table 1, results of the classification show a slight but positive impact of the annotated data on 2 data sets out of 3. BWS and IRT scores perform better on both HaSpeeDe than the baseline, namely the prediction with-

out any information from the lexicon, whereas the simple count of HurtLex occurrences results in a drop of the F1-score against the baseline. Notably, on the HaSpeeDe News test set, the additional information from HurtLex results in a lower precision but still a higher F1-score than the baseline, possibly due to the nature of newswire text which is not expected to contain many explicitly offensive terms. Performance on the AMI test set show that information provided by the lexicon always outperform the baseline. The simple HurtLex count shows a better performance than manually annotated scores, though. Such a result may be explained by the different domain addressed by AMI, which is focused on misogyny rather than HS. Furter investigation on how the perceived offensiveness of words changes within different context may be useful to better assess the impact of such resources on the classification and understanding of abusive language.

References

Basile V., Lai M., Sanguinetti M. (2018). Long-term social media data collection at the university of Turin. Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). CEUR.

Bassignana E., Basile V., Patti V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, CEUR-WS, pp 1-6.

Capozzi A.T., Lai M., Basile V., Poletto F., Sanguinetti M., Bosco, C., ... and Stranisci, M. (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the" Contro L'Odio" project. In 6th Italian Conference on Computational Linguistics, CLiC-it 2019 (Vol. 2481, pp. 1-6). CEUR-WS.

Chalmers, RP. (2012). "mirt: A Multidimensional Item Response Theory Package for the R Environment." Journal of Statistical Software, 48(6), 1–29.

De Mauro T. (2016). *Le parole per ferire*. Internazionale. 27 settembre 2016. Compiled for the "Jo Cox" Committee on intolerance, xenophobia, racism and hate phenomena, of the Italian Chamber of Deputies.

Devlin J., Chang M.W., Lee K., and Toutanova K., (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fersini E., Nozza D., and Rosso, P. (2020, December). AMI@ EVALITA2020: Automatic Misogyny Identification. In EVALITA.

Kingma D. P., and Ba J., (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kiritchenko S., Mohammad S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. vol 2, pp 465-470.

Koufakou A., Pamungkas E. W., Basile V., and Patti, V., (2020). HurtBERT: incorporating lexical features with BERT for the detection of abusive language. In Fourth Workshop on Online Abuse and Harms (pp. 34-43). Association for Computational Linguistics.

Louviere J.J and Woodworth G.G. (1991). Best-worst scaling: A model for the largest difference judgments. University of Alberta: Working Paper.
Maisto A., Pelosi S., Vietri S., Vitale P. (2017). Mining Offensive Language on Social Media. In: Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017, pp 252-256.
Poletto F., Basile V., Bosco C., Patti V., Stranisci M. (2019). Annotating Hate Speech: Three Schemes at Comparison. In: Bernardi F., Navigli R., Semeraro G. (eds) Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, CEUR-WS.org, CEUR Workshop Proceedings, vol 2481, URL http://ceur-ws.org/Vol-2481/paper56.pdf.
Polignano M., Basile P., De Gemmis M., Semeraro G., and Basile V., (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In 6th Italian Conference on Computational Linguistics, CLiC-it 2019 (Vol. 2481, pp. 1-6). CEUR.
Samejima F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph (17).
Sanguinetti M., Comandini G., Di Nuovo E., Frenda S., Stranisci M. A., Bosco C., Caselli T., Patti V., and Russo I., (2020). Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, 1-9.
Schmidt A., Wiegand M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, pp 1-10.
Tontodimamma A., (2021). Text mining for hate speech detection on social media. PhD Thesis.

Introducing Dialogic Process Analysis in Natural Language Processing

Gian Piero Turchi[1], Luisa Orrù[1], Annalisa Trovò[2], Christian Moro[1], Antonio Iudici[1], Nicolò Navarin[1]

[1]University of Padua – gianpiero.turchi@unipd.it luisa.orru@unipd.it christian.moro@unipd.it antonio.iudici@studenti.unipd.it nnavarin@math.unipd.it
[2]Pragmata Arkyreyma srl – annalisa.trovo@arkyreyma.com

Abstract
How what we say conveys a sense of reality? While content analysis considers the meaning of the text, dialogic process analysis studies the configurations of sense of reality (how the words are used), allowing to anticipate the near-future position of a speaker or a community. There is a fine line between the evolution of sense and meaning and how this sense is conveyed in a text. To what extent does the developed string capture the evolution of sense within the language process? The problem we consider does not consider the meaning, but the way the concept is expressed. What are the criteria we elaborate to give sense to what is said? How can we transfer these criteria to an artificial intelligence? The dataset we considered consists of about 170.000. We formalize the problem as a classification task, and we provide accuracy results of some neural network architectures that, combined with a set of pre-trained word embeddings, provide a good reference for the proposed dataset. The research was developed in three phases: 1. Manual analysis; 2. Automatic analysis; 3. Calculation of the margin of error of the strings. Therefore, the main goal is to make automatic identification match the operator's expertise in anticipating the use of a certain assembly of words, and what they are meant to convey within language. Analysis's output allows to recall and manage several, different and uncertain discursive configurations that can occur and not yet occurred. In other words, conveying current and future usage of words in the process of sense of reality shaping.

Keywords: Dialogic Science; MADIT methodology; NPL; machine learning; text analysis; natural language.

1. Introduction

Despite linguistics (the branch of science that studies human language) knowledge achievements on how idioms are structured in terms of grammatical and logical analysis, linguists have rarely asked themselves what could be the generative potential of a portion of speech: how does the way we say what we

say" (namely, language use) impact a person's or a community's configuration (i.e. their/its position on a certain subject)? How can we measure the impact that human interactions have on their interactants? Or how can we anticipate what will change in the speaking community when facing a decision of an administrative structure? While content analysis considers the meaning of the text (from the meaning of the single words in it), dialogic process analysis studies the configurations of meaning (i.e. how the words are used), exploiting the 24 *Discursive Repertories* (i.e. the uses of language traceable in the text, as the DNA of a text), allowing to anticipate the near-future position of a speaker or a community. Having experts defining strict logical rules to infer such properties of text (i.e. its *Discursive Repertories*) is a tedious and time-consuming task, and would result in very complex rules that have to consider all the possible scenarios in which words can be used. The application of machine learning to dialogic process analysis can alleviate these problems. We start from a set of texts, each one with the associated Discursive Repertoires labels and by using the M.A.D.I.T methodology (Methodology of Textual Computer Data Analysis; Turchi et al., 2013; Turchi and Orrù, 2014; Turchi et al., 2021). The present study seeks to determine the margin of error of a machine learning model in detecting 24 pre-determined Discursive Repertories categorised in 3 groups: Generative Repertoires (of reality configurations), Maintenance Repertoires (of reality configurations), Hybrid Repertories (they draw their generative direction from the Discursive Repertoires they combine with). Discursive Repertoires interactions that generate discursive configurations are organized in the Semi-radial Table of Discursive Repertories under their generative properties (Turchi and Orrù, 2014; Turchi et al., 2021). The table represents the main tool to guide researchers in the detection of language use and, therefore, in the measurement of a Community's inclination to change. M.A.D.I.T. moves the first one using language to assign value to a content (for example "the tsunami was difficult to handle"). When we analyse discursive processes that define Community experiences such as natural emergencies, the generative power that the text convey can be the same, even if the used words are different. For example, the text of the above example shapes the reality through a connotation of a fact (Repertory of Judgement); another text can determine a given fact (for example "nothing will ever be the same"): both discursive processes have low potential of change and tend to keep still and real the management of the event that is critical for the community. Would the machine learning have detected the same discursive processes? There is a fine line between operating a grammatical analysis and spotting the interactive process generated by a portion of speech; in fact, there is a fine line between the evolution of sense and meaning and how this sense is conveyed in a text. So, to what extent does the model developed by the researchers capture the evolution of sense within the language process? The

research was developed in three phases: 1. Manual analysis operated by the researchers: the naming of Discursive Repertories followed each time the operator's competence in anticipating the use of natural language; 2. Automatic analysis operated by machine learning, based on previously-built strings chosen by the researchers; 3. Calculation of the margin of error of the strings, which was founded on the comparison to the manual analysis. Therefore, the main goal is to make automatic identification match the operator's expertise in anticipating the use of a certain assembly of words, and what they are meant to convey within language. The higher the margin of error of the strings, the more we can anticipate they are analysing content and grammar; the lower the error, the more they are approaching language use. The Semi-radial Table of Discursive Repertories was used to determine the changing potential of portions of language (Turchi et al., 2013; Turchi and Orrù, 2014; Turchi et al., 2021). These various processes of language (Discursive Repertories) show different degrees of potential for changing the configuration of the sense of reality. Analysis's output allows to recall and manage several, different and uncertain discursive configurations that can occur and not yet occurred. That is, conveying current and future usage of words in the process of sense of meaning shaping.

2. The scientific goal

In order to name the Discursive Repertoire inside a text it is necessary to detect the rules of language used in the text. These rules are the "process properties", that have been defined for each Discursive Repertoire. These properties are the descriptions of a particular use of language. So, in order to name a Discursive Repertoire it is necessary to describe the process properties expressed by the text. What has not been done yet is to translate the process properties into a more formal language, enabling the automatic labelling of texts. This could allow to define a sentence expressed by natural language as Discursive Repertoires through an informatics support and, therefore, a formal language. Now the researchers entered a second phase. They decided to test the available tools, currently operated by the scientific Community, for the study of language use. With this data, we can train a predictive model to infer the Discursive Repertoires induced by unseen text. Now the problem is to have a predictive model, powerful enough to be able to capture the regularities in the input data (i.e. process properties). Every use of natural language generates a discursive configuration (a set of Discursive Repertoires) and dialogic process analysis does not focus on the meaning of a sentence, but on its patterns of usage. In the recent years, thanks to the advancement in optimization algorithms as well as in the available hardware, Artificial Neural Networks (NNs) have been applied to many different domains, often significantly improving the state-of-the-art in terms of predictive performance. NNs have been applied to many tasks in Natural Language

Processing, including but not limited to Question Answering, Sentiment Analysis, as well as to the task of predicting specific personality traits given a snippet of text written by a subject. Part of the success of the application of neural networks to complex data, such as text, is their ability to learn meaningful embeddings. At a word level (Mikolov, Chen, et al., 2013), it has been shown that it is possible to learn a vectorial representation for words that, given a text corpora big enough, encapsulates semantic relationships between words by means of Euclidean distance between vectors. For instance, let us define with vec("word") the function computing such a vectorial representation for the string (word) in brackets. Then vec("king") – vec("man") = vec("queen") – vec("woman"). Such embeddings are very helpful for generalizing concepts across different words. In this way, a sentence becomes a sequence of vectors (one for each word). However, the same word can have very different meaning depending on its use in the sentence. Think about the following two sentences: "Being patient has two sides to the coin" and "I am a patient of Dr. Mandela". The word "patient" is the same in the two sentences but has different meaning (perhaps etymologically connected). Then, it is possible to have different meaning, as it is possible for the same word to be used by different Discursive Repertoires, to be expressed through different rules of language (process properties). So on, the usage of the same word in different Discursive Repertoires generates different discursive configurations. For instance, we consider the word "thieves", we can say:

A "Some thieves robbed a bank during the night of january the 1st";
B "All immigrants are thieves".

Sentence A generates a discursive configuration which is a description of a fact. Sentence B is a statement that allows no opposition, things cannot be different from what is stated. In both sentences it is used the word "thieves", but they generate different discursive configuration, bringing with them a different construction of the sense of reality. For this reason, we have to construct a representation for the whole sentence, that takes in consideration the sequence of words that makes up a sentence (Quoc and Mikolov, 2014). Specific types of Neural Networks, Recurrent Neural Networks, are designed to accept as input sequences of vectors, and produce as output an embedding for the whole sequence. Finally, from this representation of the whole sentence, it is possible to name the Discursive Repertoires. This approach has already been applied to other, simpler, predictive tasks on text data (Majumder, Poria, et al., 2017). The most critical aspect of automatic language analysis, as mentioned, is the need to start from the grammatical structure. Indeed, the deductive process that takes the scientist from the portion of text to the linked generative potential conveyed by language, cannot exist without segmenting the text itself through logical analysis. How can we measure the property of generative potential, without the support

of a grammatical analysis? The research showed an operator can understand the value attached to a text regardeless logical analysis. We need to know what are the criteria we elaborate, as humans, when we listen to a portion of speech to give sense to what is said and how can we transfer these criteria to an artificial intelligence. We thus collected a set of texts, each one with the associated Discursive Repertoires labels. Note that it is not necessary for the process that generated the labeling to be formalizable, it suffices to have a set of text/label pairs. We perform three main steps: (i) data (text) collection and manual naming of Discursive Repertoires in natural language; (ii) training of machine learning models; (iii) evaluation of the error of the models.

3. Dataset

The collected texts derive from a set of researches carried out between 2017 and 2019 by the FISPPA Department of the University of Padova on behalf of AICS (Italian Cultural and Sports Association) and Ragazzi della Panchina Association. The two research programs are focused on the topic of deviance, in particular crimes of gender violence and narcotic substances consumption. The investigation group of the two research projects is composed of 5817 respondents who were asked to express their views on the subject of the specific research investigation responding to 20 open questions. The total words in the text are 69,271 which are organized in 3741 Maintenance Discursive Repertoires, 983 Hybrid and 1093 Generative Discursive Repertoires. A second set of texts was collected from Italian newspaper articles about the condition of savers following the global economic crisis. This second set of texts contains 99,870 occurrences, divided in 2649 excerpts of text, of which 1,390 Maintenance Repertoires, 663 Hybrid Repertoires and 596 Generative Repertoires. To make the task simpler for machine learning algorithms, we balanced the dataset under-sampling the majority class (Maintenance Repertories). Our final dataset has 1800 Maintenance Repertoires, 1646 Hybrid Repertoires and 1689 Generative Repertoires. In this first work, we consider a simplified classification task in which we group the 24 Repertoires in the 3 macro-classes (or groups) described in Section 2, obtaining the statistics reported above. The resulting task is simpler compared to the classification of the 24 Repertoires, and given the relatively small amount of data available, we decided to focus on this simpler setting.

4. Experiments

In this paper, we focused on deep learning models to present strong baselines on the proposed dataset. Concerning NLP tasks, relatively recent research has shown that word embeddings, i.e. representing each word with a general-purpose fixed-size vector that encodes the semantic similarity among words, is very beneficial in many settings. We thus used the Italian word embeddings released with

(Cimino, De Mattei, et al., 2018). For the task of text classification, two main models are available: Recurrent Neural Networks, specifically Long Short-Term Memory and 1-dimensional Convolutional Neural Networks. In the next section, we report experimental results of both models in the proposed dataset. We report results on the balanced versione of the dataset. We divided the data in training, validation and test folds with, respectively, 55%, 20% and 25% of the data. We considered Recurrent (LSTM) models, and 1-dimensional Convolutional Neural Networks (1D-CNN) models. We tested LSTM models with one or two LSTM layers, followed by a fully connected layer and the output layer, with ReLU and softmax activation function, respectively. For 1D-CNN, after preliminary results we considered two convolutional layers, with filter size of each layer in {2, 3, 4}. After each convolutional layer, a global max pooling operator is applied. The outputs of all such layers is concatenates and fed to the output layer with a softmax activation function. We also consider a model that computes at the same time the filters from size 2 to 4, and concatenates their output (inspired by the Inception model). We validated the number of neurons in {32, 64} for LSTM and in {50, 100, 150, 200} for 1D-CNN. Since the models we tested are small, their behavior is pretty stable.

4.1. Results

Our experimental results are reported in Table 1. We recall that, for a balanced 3-classes problem, the accuracy of a random classifier would be 0.33. The predictive performance of all the considered models are above 0.53, meaning that they are actually able to capture some signal in the data. LSTM results to be the best performing model. As for 1D-CNNs, using multiple single filter sizes at the same time seems to help the model to achieve similar performance to LSTM. This is reasonable since we are not forcing a fixed size for the learnt features. We recall that these results are meant to be baseline references. Our results suggest that existing Deep Learning models seem to capture some signal in the data, but more research is needed to define effective models for this new task. In the future, we plan to extend our approach adopting state-of-the-art BERT-based word embeddings for the Italian language such as (Polignano, Basile, et al., 2019; or https://github.com/musixmatchresearch/umberto).

Table 1: Experimental results of different deep learning models on the proposed dialogic process analysis dataset.

Model	Hidden neurons	Filter size	Accuracy
LSTM	32-32	-	0.554
1D-CNN	150-300	2	0.545
1D-CNN	150-300	3	0.534
1D-CNN	150-300	4	0.543
1D-CNN	150-300	2-3-4	0.550

5. Conclusions

The scientific questions that moved this paper are: how can we transfer the same logic to language measurement? How can we measure the property of generative potential of text, without the support of a grammatical analysis? The research showed an operator can do so. Thus, first and foremost, the question we need to answer is: what are the criteria we elaborate, as humans, when we listen to a portion of speech to give sense to what is said? And, secondly, how can we transfer these criteria to an artificial intelligence? Answering these questions would have an impact on our Community far more universal then increasing the efficacy of events management such as, for instance, make an important decision for people health. It would effect all areas of knowledge, because language "invades" any situation or scenario we know. It is the one necessary and sufficient condition to establish a sense of reality; thus, everything we know is determined by the language process used to know it. In this paper we proposed a dataset that models dialogic process analysis, a problem that is orthogonal to many other tasks in literature, and we believe it may be helpful in the future for the definition of more advanced NLP models and for generating more expressive word and sentence embeddings. Learning to understand and manage language's uncertain flow is surely the first step; becoming so precise we can render these skills automatic, would be the next big revolution.

References

Cimino A., De Mattei L., and Dell'Orletta F. (2018). Multi-task learning in deep neural networks at EVALITA 2018. *CEUR Workshop Proceedings*, 2263.

Dos Santos C. N. and Gatti M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014: Technical Papers*, pages 69–78.

Majumder N., Poria S., Gelbukh A., and Cambria E. (2017). Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, 32(2):74–79.

Mikolov T., Chen K., Corrado G., and Dean J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 1–9.

Pennington J., Socher R., and Manning C.D. (2014): GloVe Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Polignano M., Basile P., de Gemmis M., Semeraro G., & Basile V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. *CEUR Workshop Proceedings*, 2481.

Quoc V. Le and Mikolov T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning, PMLR*, volume 32, pages 1188 – 1196.

Szegedy C., Wei Liu, Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Turchi G.P., Dalla Riva M.S., Ciloni C., Moro C., and Orrù L. (2021). The Interactive management of the SAR-CoV-2 Virus: The Social Cohesion Index, a Methodological Operational Proposal. *Frontiers in Psychology*, 12, 559842. doi: 10.3389/fpsyg.2021.559842

Turchi G.P., Girardi A., Pinto E., and Orrù L. (2013). Efficacy measurement in penitentiary treatment: a methodology to meet the requirements of current legislation. In *International Congress of the Advanced High School of Criminological Sciences - CRINVE*.

Turchi G.P. and Orrù L. (2014). *Metodologia per l'analisi dei dati informatizzati - M.A.D.I.T. Fondamenti di teoria della misura per la scienza dialogica*. Napoli: Edises.

Turchi G.P., Romanelli M. (2019). Dialogical mediation as an instrument to promote health and social cohesion: Results and directions. *Comunicação e Sociedade*, special vol., 2019, pp. 131 – 140.

Turchi G.P., Romanelli M., Bonazza F., and Girardi A. (2013). Discursive Repertory, Discursive Configuration. In *Encyclopedia of Critical Psychology*, chapter 569/570. Springer.

Convolution et marqueurs multidimensionnels. Description des représentations genrées dans un corpus de films français

Laurent Vanni, Magali Guaresi, Véronique Magri

Université Côte d'Azur, CNRS, BCL - laurent.vanni@univ-cotedazur.fr
magali.guaresi@univ-cotedazur.fr veronique.magri@univ-cotedazur.fr

Abstract
Convolutional neural networks allow new representations of texts that extend the standard statistical approaches. By combining frequency and context of words as well as allowing multidimensional treatments (graphical form, lemma and part of speech), convolution leads to the extraction of *motifs*, i.e. complex linguistic patterns that are likely to feed interpretation. In this paper, this architecture is tested on movie scripts in order to explore the hypothesis of a gendered differentiation of female and male dialogues.

Keywords: Gender, Movies, Convolution, *Motifs,*Textual Statistic Analysis, Deep learning.

Résumé
Les réseaux convolutionnels appliqués aux textes permettent de nouvelles représentations du texte, prolongeant les résultats de l'analyse de données textuelles. En combinant approche fréquentielle et séquentielle des textes et en permettant un traitement multidimensionnel des mots (forme graphique, lemme et code grammatical), la convolution aboutit à l'extraction de motifs, c'est à dire de patrons linguistiques complexes susceptibles d'alimenter l'interprétation. Dans ce papier, ce type de modèle est mis à l'épreuve d'un corpus de scripts cinématographiques dans le but d'explorer l'hypothèse d'une différenciation genrée des dialogues de femmes et d'hommes.

Mots clés : Genre, Cinéma, Convolution, Motifs, Statistique textuelle, Deep learning.

1. Introduction

Les réseaux convolutionnels appliqués aux textes (Collobert et Weston 2008) permettent de dépasser l'approche fréquentielle de la statistique classique, en croisant l'axe paradigmatique (la sélection des unités linguistiques) et l'axe syntagmatique (la combinaison de ces unités dans la linéarité du passage)[1]. Par

ailleurs, comme en analyse de données textuelles (ADT), la convolution est capable de considérer plusieurs représentations des mots : forme graphique, catégorie grammaticale et lemme. Ces représentations multidimensionnelles (ou multi-*channels*) permettent, non seulement, d'obtenir de meilleures performances en termes de classification (prédiction), mais aussi d'extraire des motifs au sens de (Mellet et Longrée 2009), susceptibles d'alimenter la description et l'interprétation des textes (Mayaffre et Vanni 2021). À partir d'un corpus de scripts de films français contemporains, cette étude propose une description des dialogues en fonction du sexe des locuteurs/locutrices et du sexe des cinéastes. Nous montrerons que l'approche convolutionnelle multicritère prolonge l'ADT classique en mettant en évidence des patterns linguistiques profonds - combinant les activations des trois *channels* - qui interroge d'une part le deep learning et sa capacité à extraire de l'information linguistique complexe et d'autre part les représentations de genre - compris comme un système de bicatégorisation des sexes et des représentations qui y sont associées (Achin et Bargel, 2013) au cinéma.

2. Méthode et corpus

Les réseaux convolutionnels, initialement conçus pour l'analyse d'images, utilisent souvent plusieurs channels, c'est-à-dire plusieurs tableaux de données différents, pour représenter une même image en entrée de plusieurs façons. La méthode la plus répandue consiste à utiliser le codage RGB (Red, Blue, Green) pour représenter l'image suivant trois nuances de couleurs. L'idée de cet algorithme est de permettre au modèle d'extraire des contrastes particuliers qui utilisent des couleurs différenciées. Pour les textes, il est possible d'appliquer le même raisonnement pour encoder séparément les mots avec leur forme graphique, leur catégorie grammaticale et leur lemme. Cette approche présente deux avantages majeurs : d'une part, elle autorise le modèle à mixer syntaxe et lexique pour distinguer deux classes différentes dont elle améliore potentiellement la précision, d'autre part, elle permet de restituer des motifs linguistiques profonds basés sur le calcul du Text Deconvolution Saliency (TDS) (Vanni et al. 2018), implémenté dans Hyperbase[2]. La capture de ces motifs est une méthode exploratoire fondée sur l'analyse des zones d'activations les plus fortes (pic d'activation du TDS) dans le texte. Ces passages que le deep learning souligne pour prendre une décision (classification) ont vocation à interroger l'analyste sur

[1] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

[2] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

les données, leur nature et les contrastes qui discriminent chaque texte (Vanni 2021). C'est précisément cette potentielle exploration fine qui nous pousse à utiliser cette méthode dans le cadre d'une étude des représentations genrées dans les films. Celles-ci seront décrites par l'analyse des spécificités (section 3.1), prolongée par l'exploration des motifs profonds qui ouvre des parcours de lecture nouveaux et expose des résultats moins attendus (section 3.2).

Le corpus que nous étudions ici est composé de 86 films français, issus de la base de scripts en ligne Lecteurs anonymes[3]. Des travaux ont montré qu'il était possible d'extraire de manière automatique les répliques d'un corpus de scripts et de les attribuer à un sexe (Apoorv et al., 2015). Ici, nous avons utilisé une approche similaire en nous fondant sur un dictionnaire de noms propres et un contrôle manuel. Pour nos expérimentations, le corpus est divisé en quatre sous-corpus regroupant les répliques selon le sexe des personnages et le sexe des réalisateur.rices.

Tableau 1. Description du corpus « Scripts »

	Occurrences	Vocabulaire	Hapax
Personnages féminins par Réalisatrices	82899	6421	3508
Personnages féminins par Réalisateurs	214503	12821	6953
Personnages masculins par Réalisatrices	49736	5523	3264
Personnages masculins par Réalisateurs	305536	16483	8595

La partition du corpus est de taille inégale, les films de réalisatrices et les dialogues de personnages féminins étant sous-représentés. Mais loin d'être un biais de corpus, cette répartition traduit la réalité de la production cinématographique française. D'après une étude publiée par le Centre national de la cinématographie en 2019, seuls 27% des fims sortis en 2017 ont été réalisés par des femmes et les personnages féminins n'y ont qu'un tiers du temps de parole.

Le modèle entrainé pour cette étude utilise 80% du corpus pour l'entrainement et 20% pour la validation. Les classes choisies correspondent aux quatre parties du corpus décrites dans le tableau 1 et la précision du modèle sur ces classes atteint 85% sur l'ensemble de validation.

[3] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

3. Etude de cas : des spécificités aux motifs genrés dans un corpus de scripts français

3.1. Spécificités

L'analyse de données textuelles rend compte, par le calcul des spécificités (Lafon 1980), des mots sur-utilisés[4] dans les répliques des personnages féminins comparées à celles des personnages masculins. Qu'ils soient produits par des réalisatrices ou des réalisateurs, les dialogues construisent des protagonistes stéréotypiquement genrés qui n'emploient pas les mêmes mots selon leur sexe. D'un point de vue thématique d'abord : les femmes évoquent, dans les films de réalisatrices, des sujets réputés féminins comme celui de la reproduction (« pilule » [14,52], « enceinte » [10,79], « grossesse » [8,95], etc.)[5], là où les hommes, en particulier dans les films de réalisateurs, abordent de larges thèmes publics, « politiques » [4,6] ou économiques (« entreprise » [3,86], « société » [3,73], « justice » [3,57]), reproduisant ainsi la traditionnelle dichotomie féminin-privé versus masculin-public. La différenciation genrée s'opère également pour les personnages masculins des réalisatrices présentés comme compétitifs (« concours » [6,17]) voire sportifs (« ballon » [4,7]), ou encore bricoleurs (« démonter » [4,3]). Les mots caractéristiques des femmes dans les films de réalisateurs montrent, en revanche, des personnages élaborés dans le cadre des relations familiales (« maman » [3,35]) ou affectives (« chéri » [4,35]). Les spécificités lexicales confirment l'existence de différences sensibles entre les discours des femmes et des hommes au cinéma, leur construisant ainsi des places et des rôles différents (Brey 2020).

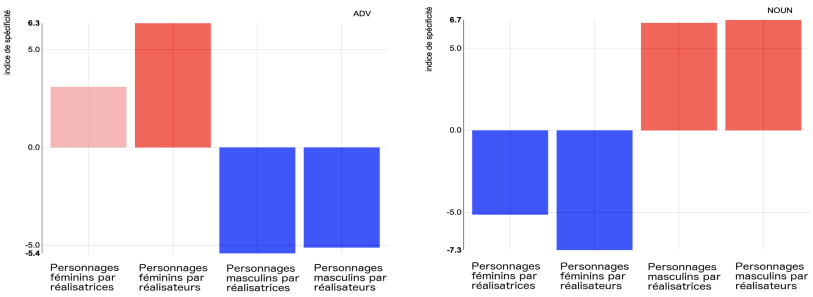


Figure 1. Sur-utilisation des adverbes et sous-utilisation des noms dans les répliques de personnages féminins

[4] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

[5] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

Pour dépasser la simple analyse thématique, la distribution statistique des catégories grammaticales suggère des différences syntaxiques entre les répliques prononcées par les femmes et par les hommes (figure 2). Les premières sur-utilisent les adverbes et les pronoms alors que les hommes emploient davantage de noms, laissant penser à deux économies discursives distinctes (discours verbal et modalisé des femmes versus discours nominal et substantiel des hommes).
Si des méthodes statistiques poussées nous autoriseraient une analyse plus fine, utilisant par exemple les segments répétées (Salem, 1986) ou les (poly)cooccurrences spécifiques (Vanni, 2016), l'extraction des motifs profonds par le deep learning offre un parcours interprétatif complémentaire qui croise à la fois le lexique et la syntaxe dans une même analyse.

3.2. De l'urne au réseau : les motifs genrés profonds

Nous l'avons vu, le deep learning offre une lecture nouvelle des données. À première vue, les zon es d'activation (TDS) confirment les analyses préliminaires

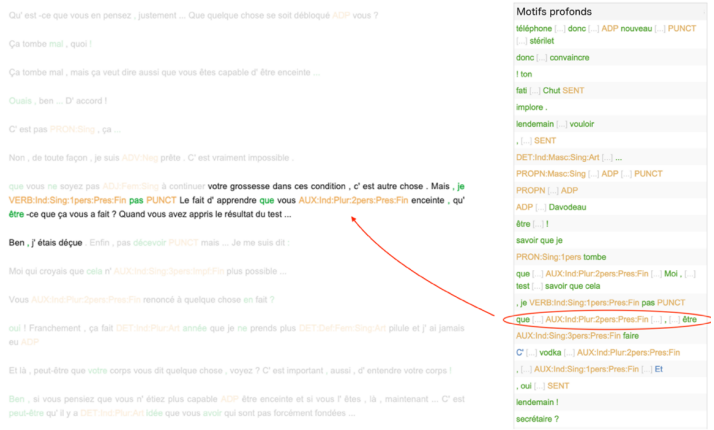


*Figure 2. Sur-utilisation et activation du motif « que *** AUX *** LEM:être »[1] dans les répliques de femmes chez les réalisatrices*

faites avec l'ADT classique.
Le lexique occupe une place importante dans les marqueurs qui permettent de discriminer chacune des classes. Ainsi, nous retrouvons, pour la classe « Personnages féminins par réalisatrices » (Figure 3), le mot « stérilet » qui marque la plus forte activation du réseau. D'autres unités comme l'adverbe « donc » ou le point d'exclamation, présents dans les passages-clés, sont aussi des marqueurs attestés par le calcul de spécificités. Mais la plus-value de la méthode ne se situe pas dans ces premiers marqueurs détectés. Un peu plus loin dans la liste des motifs (20 premiers résultats), un premier schéma linguistique semble

proposer un véritable motif (Figure 3). Composé du subordonnant « que » suivi immédiatement ou non d'un auxiliaire[6] conjugué à la deuxième personne de pluriel, et plus loin, du verbe « être », ce passage peut, en effet, être considéré comme un motif profond (Vanni, 2021) en tant que suite de marqueurs non contigus et multidimensionnels qui marque une singularité dans les dialogues de femmes chez les réalisatrices et que la statistique classique peine à détecter automatiquement.

« Que » est un terme polysémique qui recouvre en langue plusieurs natures grammaticales. Dans le corpus, sous la multiplicité des réalisations grammaticales, des traits communs émergent. Dans l'exemple figurant sur la figure 3, le motif renvoie à une interrogation adressée par une conseillère du Planning familial à une jeune femme. Celle-ci lui demande : « Le fait d'apprendre que vous étiez enceinte, qu'est-ce que ça vous a fait ?[7] ». Outre le lexique repéré par la statistique comme « féminin », la structure « que + auxiliaire + *** + être », repéré par l'algorithme de deep learning comme caractéristique des prises de parole féminines dans les films de réalisatrices, montre une tendance à la mise au premier plan de la perception affective ou évaluative des faits.

Dans la majorité des motifs tirés du corpus, ces processus de retours sur les faits se concrétisent par un retour sur un « dire », au travers de formes de la parole rapportée ou de la reformulation :

« Donc, vous me disiez que vous avez vu le médecin hier, en urgence… C'était tard non ?[8] »

« ça veut dire que vous êtes capable d'être mère vous aussi[9] »

La complétive introduite par « que » est placée sous la rection du verbe de parole « dit », soit pour reprendre les propos d'un tiers dans une forme de discours indirect (premiers exemple), soit pour reformuler et utiliser la formule « vouloir dire » qui introduit une glose explicative (deuxième exemple). Les tournures emphatiques sont encore plus remarquables quand elles thématisent le prédicat, c'est-à-dire l'évaluation du fait, en le plaçant en début d'énoncé. L'emphase repose dans l'extrait suivant sur la redondance du sujet exprimé sous la forme développée de la complétive.

« ça ne m'étonne pas que vous soyez dehors, vous aimez bien ce temps-là. C'est

[6] Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE38-0012-01 et du Programme d'Investissement d'Avenir UCA JEDI ANR-15-IDEX-0001.

[7] *Les Bureaux de Dieu*, 2008.

[8] *Les Bureaux de Dieu*, 2008.

[9] *Les Bureaux de Dieu*, 2008.

magnifique votre site[10] »
Les répliques des femmes dans ce contexte paraissent manifester une dissociation marquée entre le *dictum* (ce qui est dit) et le *modus* (l'attitude exprimée par l'auteur.e de l'énoncé) en faveur de la mise en évidence du second placé volontiers au premier plan dans la phrase, en position frontale. L'énoncé manifeste un doublage modalisateur persistant.

4. Conclusion

Cette étude montre, sur le plan méthodologique, que les réseaux convolutionnels offrent aujourd'hui de nouvelles représentations des textes. En combinant l'approche fréquentielle et séquentielle et en considérant ensemble les activations des trois *channels* (lexique, lemme, code grammatical), le modèle permet de mettre au jour des motifs linguistiques profonds susceptibles de prolonger les résultats en statistique textuelle sur les représentations discursives de genre au cinéma. Ces motifs présentent une double plus-value : d'une part, ils sont des unités contextuelles *de facto* porteuses de sens et d'autre part, ils combinent différents niveaux de granularité de la textualité pour être de nouveaux observables complexes du genre.
Ces premières analyses, à mettre à l'épreuve d'autres données filmiques, suggèrent la prégnance des stéréotypes genrés dans l'élaboration des personnages féminins et masculins. Outre les spécialisations lexicales, qui cantonnent les femmes au privé et à la reproduction, les outils de l'analyse de données textuelles et du deep learning montrent des économies discursives sexuées différentes : aux hommes, le discours nominal, référentiel ou informatif ; aux femmes le discours adverbialisé et modalisé. La répartition de la parole entre les sexes dans les films est régie par des rapports de pouvoir, proches de ceux observés dans les conversations dans nos sociétés (Monnet 1998). Les répliques féminines sont moins fréquentes, plus courtes et souvent sous forme de questions ou de demande de précisions sur ce qui vient d'être dit. Comme l'exemplifie le motif étudié dans la dernière section, les répliques féminines se démarquent par une mise au premier plan de l'énonciation avant le contenu même de l'énoncé. Ce marquage plus fort de la subjectivité des énonciatrices dans les scénarios traduit moins une emprise directe et active sur le monde qu'un retour, évaluateur ou affectif, par rapport aux évènements qui sont évoqués et rapportés dans les dialogues.

[10] Roxane, 2019. Dans cet exemple, le verbe « être » est au subjonctif présent. L'embedding (word2vec) utilisé par le modèle fait apparaître cette catégorie dans les 10 plus proches voisins (cosine distance) de l'auxiliaire à l'indicatif. Nous faisons donc l'hypothèse, ici, que cette variation fait partie des motifs soulignés par le modèle. Cette hypothèse sera validée dans des travaux ultérieurs.

Références

Achin C. et Bargel L. (2013). *Dictionnaire Genre et Science politique*, Paris, Presses de science politique.

Apoorv et al. (2015). Keys Female Characters in Film Have More To Talk About Beside Men : Automating Betchdel Test, *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 830–840, Denver, Colorado.

Brey I. (2020). *Le Regard féminin. Une révolution à l'écran*, Paris, Edition de l'Olivier.

Collobert, R. et J. Weston (2008). A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning. *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland : ACM, 160–167.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots,* 1: 127–165.

Lécossais S. (2017). Les mots de la grossesse. *Études de communication*, 48 : 155-176.

Mayaffre D. et Vanni L. (2021). *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Champion, Paris, 2021.

Mellet S. et Longrée D. (2009). Syntactical motifs and textual structures. *Belgian Journal of Linguistics*. T. 23 : 161–173.

Monnet C. (1998), La répartition des tâches entre les femmes et les hommes dans le travail de conversation », *Nouvelles Questions Féministes*, vol. 19 (n°1) : 9-34.

Salem A. (1986). Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, vol 1 (n°2) : 5-28.

Vanni L. (2021). *De l'analyse statistique de données textuelles aux réseaux de neurones artificiels. Vers des motifs linguistiques profonds*. Thèse de doctorat, Université Côte d'Azur.

Vanni L., Ducoffe M., Mayaffre D., Precioso F., Longrée D., et al. (2018). Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis. *In : 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 548–557, Melbourne, Australia.

Vanni L. et Mittmann A. (2016). Cooccurrences spécifiques et représentations graphiques, le nouveau "Thème" d'Hyperbase. *12es Journées internationales d'Analyse statistique des*

Crise du Covid et ressenti des acteurs du secteur social et médico-social : une analyse textométrique de journaux de bord et d'entretiens

Pierre Wavresky

CESAER UMR1041, INRAE, Institut Agro, UBFC –
pierre.wavresky@inrae.fr

Abstract
73 diaries and interviews from professionals and beneficiaries of the social and medicosocial sector during the 1st covid wave were analysed using textometry tools, highlighting 15 types of discourse, sometimes focused on a feeling, such as fear, and sometimes focused on a specific group of people, such as children or the elderly. An analysis of the feeling of fear showed that its strong presence at the beginning of the diary gradually gave place to the feeling of happiness/unhappiness.
Keywords: textometry, clustering, personal pronoun, Covid-19, social and medicosocial sector.

Résumé
73 journaux de bord et entretiens émanant de professionnels et d'usagers du secteur social et médico-social durant la 1ère vague du covid ont été analysés grâce aux outils de la textométrie, mettant en lumière 15 types de discours, parfois centrés sur un ressenti, comme la peur, parfois plutôt centrés sur un public, comme les enfants ou les personnes âgées. Une analyse du sentiment de peur a montré que sa forte présence en début de journal a cédé progressivement la place au sentiment de bonheur/malheur.

Mots clés: textométrie, classification, pronoms personnels, Covid-19, secteur social et médico-social.

1. Introduction

La crise sanitaire due à la pandémie de Covid19 a eu un impact particulièrement sensible sur les personnes vulnérables (résidents de Ehpad, mineurs handicapés, personnes en foyer d'hébergement…) et les professionnels du secteur social ou médico-social les accompagnant. Tout d'abord par la maladie et la mort bien sûr, mais aussi par le changement des relations entre les différents acteurs (professionnels, personnes vulnérables, autres personnes) du fait du risque de contagion. Les conditions de travail des acteurs en ont été fortement affectées.
La *Fédération des Centres Régionaux d'Étude d'Actions et d'Information, en*

faveur des personnes en situation de vulnérabilité (ANCREAI) a mené et recueilli 73 témoignages (essentiellement des journaux de bord mais aussi quelques entretiens ou séries d'entretiens) émanant pour la plupart de professionnels, mais aussi de responsables d'associations et de personnes accompagnées tout au long de la première vague de Covid19. La diversité des profils des témoins a été recherchée afin d'obtenir un vaste panorama du secteur. Cette diversité touche différentes caractéristiques des témoins : leur profession (infirmière, psychologue, éducateur spécialisé, agent d'entretien, enseignant spécialisé, directeur de centre …), le type d'établissement dans lequel ils travaillent, (*Esat*, *Ehpad*, foyer d'hébergement, association de précarité, *Csapa*, *Service d'Éducation Spécialisée et de Soin À Domicile* …) et sa localisation (9 régions situées en zone verte ou rouge) ainsi que le public dont ils s'occupent (addictologie, majeurs protégés, polyhandicapés, personnes âgées…). Le témoignage de 5 résidents d'un foyer d'hébergement a permis de contraster leur discours avec celui des professionnels travaillant dans le même type d'établissement : quels thèmes partagent-ils, quels thèmes sont spécifiques aux résidents ou à ces professionnels ?

Les journaux ont été rédigés au fur et à mesure par les témoins, selon des consignes d'écriture volontairement limitées, « racontez la crise générée par l'épidémie dans la structure où vous travaillez », afin de ne pas brider leur expression. Les entretiens semi-directifs, quant à eux, avaient comme thème général la façon dont s'est passé le confinement et le déconfinement, pour l'interviewé, le service et les personnes aidées. Ils ont été menés par deux professionnels de l'ANCREAI. Journaux et entretiens seront appelés « témoignages », dans la suite du texte.

Pour faire cette analyse, menée dans le cadre de *l'ANR « CRISEsoc.JdB - Les acteurs du social et du médico-social aux prises avec la crise générée par la Covid-19 »*, nous avons mobilisé les outils de la statistique et de la textométrie. Les traitements informatiques et statistiques ont été faits grâce aux logiciels R (R Development Core Team, 2005) et Iramuteq (Ratinaud, 2009).

Les relation entre les acteurs étant un des thèmes analysés, nous avons, dans un premier temps, explicité le codage des noms des personnes et des pronoms correspondants, selon le statut de la personne.

Puis nous avons mené une analyse textométrique des écrits, tout d'abord par le comptage des occurrences des lemmes, déjà riche d'enseignements. Afin de tenir compte du contexte, nous avons ensuite fait une classification descendante hiérarchique (Reinert, 1973) mettant en évidence 15 types de discours, répartis en trois grands groupes.

Enfin nous avons étudié l'évolution du traitement des thèmes de la peur, du dépit et du bonheur au sein de chaque écrit, afin de voir s'il y avait noircissement du discours.

2. Codage du nom des personnes et des pronoms personnels

Un des objectifs de l'analyse était de voir, dans les différents discours tenus, comment sont désignées les personnes aidées, ainsi que les collègues. La première question est de savoir si le scripteur parle d'eux comme d'un groupe, auquel cas il utilisera des pronoms pluriels, ou comme de personnes particulières, le pronom singulier sera alors utilisé. La deuxième question est de savoir si la personne dont on parle est « active » ou si c'est une personne à laquelle on s'adresse, dans le premier cas un pronom sujet (il, elle, ils, elles) sera utilisé, sinon ce sera un pronom complément d'objet direct (le, la, l, les) ou indirect (lui, leur).

Il est nécessaire dans un premier temps de coder les personnes selon qu'elles sont des personnes aidées, des personnes qui sont des proches du/de la scripteurice (famille et amis), des personnes qui sont des collègues (ou leur famille, leurs amis).

Les pronoms personnels (il, ils, elle, elles, le, la, l, les, lui, leur) sont codés ainsi:

. un tiret bas : ce sont des personnes aidées (_il, _ils, _elle, _elles, _le…)

. deux tirets bas : ce sont les proches de la/du scripteurice (__il, __ils, __elle…)

. trois tirets bas : ce sont les collègues et leurs proches (___il, ___ils, ___elle...)

Dans les textes anonymisés, il est intéressant de savoir si la personne dont on parle est une personne aidée, un collègue (ou un proche de collègue), une autre personne. Le nom est en 3 parties, séparée par un tiret bas. La première est le nom anonymisé lui-même. La deuxième est la qualité de la personne (« p » pour personne aidée, « s » pour un ou une collègue du personnel médical ou médico-social ou social, « a » pour les autres). Ainsi le Professeur Raoult est considéré comme « a », car il n'est pas un collègue direct. La frontière est parfois délicate à tracer dans des cas similaires. La troisième est le genre de la personne (f pour une femme, m pour un homme, i quand nous n'avons pu, d'après le contexte, trancher). Ainsi *_sophie_s_f* est une collègue (*s*), qui s'appelle *Sophie* et est une femme (*f*), *_o_p_i* est une personne aidée (*p*), qui s'appelle *o*, et on ne sait pas si c'est un homme ou une femme (*i*).

3. Analyse statistique et textométrique des témoignages

La longueur moyenne d'un témoignage est de 5362 mots, soit une dizaine de pages ; la moitié compte entre 1835 et 7742 mots. Une entrée d'un témoignage, qui peut être un jour ou une semaine, comporte en moyenne 133 mots, et la moitié des entrées se situe entre 13 et 156 mots.

Nous allons tout d'abord compter les occurrences des lemmes dans le corpus, ce qui donnera un premier aperçu des thématiques présentes dans les témoignages, puis nous mettrons en évidence 15 types de discours fortement présents au sein des témoignages.

3.1. Occurrences

Le comptage des lemmes met en lumière l'importance d'un 1er thème, celui de la temporalité, et ce sous différents aspects. Le lemme le plus courant est *aller*, et il est utilisé 74% des fois, en utilisant une fenêtre kwic de 5 mots autour des formes du verbe *aller* au présent et à l'imparfait, sous la forme d'un futur immédiat (« l'internat ne *va* pas *rouvrir* », Auteur 46), indiquant parfois une certaine incertitude. Le 2ème lemme le plus courant est *temps*, avec des locutions comme « prendre le temps », « en même temps », et très rarement dans l'acception météorologique. Les lemmes traduisant une temporalité plus précise, comme *semaine*, *jour*, *journée*, *h* (indiquant un horaire), *heure* (indiquant une durée) sont très bien représentés aussi.

Le 2ème thème important est les gens, que ce soient des personnes aidées (*famille*, *personne*, *enfant*, *_il*, *_ils*, *résident*…) ou les collègues (*collègue*, *professionnel*, *___elle*, *service*, *équipe*…). Les proches des scripteurs sont moins souvent cités mais seront toutefois très présents dans la classe 2 (cf. infra, graphique 1).

Un 3ème thème est le travail (*travail*, *réunion*, *travailler*, *activité*…) et son organisation durant la pandémie (*télétravail*, *mail*, *téléphonique*…).

3.2. Les 15 classes de discours

Nous avons 3 grands groupes de classes (cf. infra, graphique 1), tout d'abord un groupe constitué d'une seule classe (15) sur l'addiction, puis un groupe constitué de 8 classes (de la classe 2 à 14 sur le graphique) sur le ressenti des gens (bonheur, peur…) et les contraintes (horaires, mesures de protection), et enfin un groupe de 6 classes sur le travail et son organisation, parfois centré sur un public particulier.

Dans la classe 15 (addictions), différents produits (*alcool*, *cannabis*, *cocaïne, produit*…) sont cités, parfois dans la même phrase, ainsi que les produits de substitution (*produit*, *méthadone*…). Le problème de la précarité est également abordé dans cette classe (*manche*, *abri*, *insérer*, *rue*…). La façon de parler des gens est spécifique car très souvent au pluriel (*personnes*, *_ils*, *patients*…), cette généralisation étant notamment due au fait que les données sont particulièrement sensibles et que l'interviewé ne nommera jamais de personnes particulières souffrant d'addiction, tout au plus l'appellera-t-il « *_il* ». Une autre raison potentielle de ce suremploi du pluriel est que l'interview (car les professionnels travaillant dans ce secteur ont été interviewés, principalement) incite plus à la généralisation qu'un journal proprement dit.

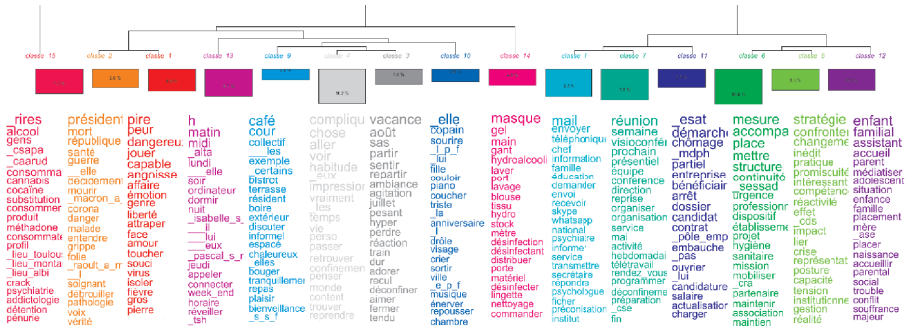


Figure 1. Les 15 classes de discours

La classe 2 traite notamment du discours du *président* de la *république* *_macron_a_m* sur la *guerre*. Le thème de la maladie et de la mort est surreprésenté (*_Covid19*, *corona*, *décéder*, *décès*, *malade*…), avec un lien fort avec les proches du scripteur (*__elle*, *__ils*, *__lui, __enfants, __fille*…). Le discours est empreint d'incertitude, qui s'exprime notamment par un suremploi des verbes au conditionnel, dans la classe 1 que nous allons voir mais surtout dans celle-ci.

Le discours de la classe 1, spécifique des témoignages des structures dédiées à un public adulte, est celui de la peur : *peur*, *dangereux*, *angoisse*, *émotion*, *insécurité*, *inquiétant*, *stress*… Trois Tgen (Types Généralisés), rassemblant des lemmes autour de trois thèmes (la Peur, le Bonheur/Malheur, le Dépit) permettent de caractériser les classes et de voir s'il y a un noircissement du discours tout au long de la période. Ainsi le Tgen Peur est spécifique de la classe 2, mais surtout de la classe 1.

La classe 13 est celle des horaires dans toutes ses composantes (*h*, *horaire*, *heure*, *matin*, *midi*, *soir*, *nuit*, *dormir*, *lundi*, *jeudi*, *mardi*, *dimanche*, *journée*…) et des collègues (*___elle*, *___lui*, *___eux*, *denis_s_m*…) parfois aussi des proches (*__mari*…) car les horaires ont un impact sur la vie familiale (« *__mari* un peu *déçu* par rapport à l'étendue du *planning* », Auteur 31). Ce discours est spécifique de la 2ème période (entre le 1er quartile et la médiane de date, voir le point 4 pour la définition des périodes), ce qui signifie que les difficultés qu'il traduit ne se rencontrent ou ne s'expriment pas dès le début des témoignages.

La classe 9 est celle du plaisir (*café*, *boire*, *plaisir*…), voire du bonheur (*bonheur*, *heureux*), d'une bonne atmosphère (*chaleureux*, *tranquillement*, *bienveillance*), obscurcis par les contraintes (*espacé*, *restreindre*, *interdire*).

La classe 4 est celle des contraintes du confinement et des incertitudes du déconfinement (*compliqué*, *difficile*, *déstabilisant*) qui pèsent sur les personnes aidées (*_ils*, *_les*, *_leur*).

La classe 3 est celle des vacances (*vacances*, *partir*, *repartir*, *juillet*, *août*…), ce discours est spécifique des résidents (qui sont tous dans le foyer d'hébergement d'Alta).
La classe 10 est celle des personnes âgées, parfois des enfants aussi, et d'émotions diverses (*content*, *sourire*, *souriant*, *triste*, *pleurer*, *marre*…), ce sont des histoires particulières et non générales (*_e_p_f*, *_l_p_f*, *_elle*, *_lui*, *_la*, *_le*, *copain*, *fille*, *ami*…). Les lemmes des émotions ne sont pas extrêmes : peu de *mal_être*, *bonheur*… mais plutôt *plaisir*, *content*, *pleurer*… Ces lemmes concernent aussi bien le scripteur que les personnes aidées (« j'ai *plaisir* à voir cet enfant et sa maman lors de la visite » Auteur 62).
La classe 14, un peu à part, se penche sur les produits et équipements de protection (*porter*, *masque*, *hydroalcoolique*), les précautions (*laver*, *lavage*, *main*, *désinfection* [de la] *porte*, *distance* [d'un] *mètre*…) et des problèmes d'approvisionnement (*stock*, *distribuer*, *commander*…).
La classe 8, fortement représentée dans le discours des professionnels travaillant pour un public d'enfants en situation de handicap sensoriel, concerne les relations à distance (*transmettre*, *envoyer*, *mail*, *skype*…) reliant l'ensemble des acteurs (personnel vers le personnel, vers la direction, les familles, familles vers le personnel…). Avec la classe 7, qui concerne plus les réunions liées au confinement et à la reprise (*réunion*, *équipe*, *présentiel*, *mai*, *reprendre*, *déconfinement*…), ce discours des contacts et de leur modification totalise presque 20% du corpus.
La classe 11 est centrée sur le public de l'insertion par le travail (*_esat*, *chômage*, *candidat*…), avec une forte dimension administrative (*dossier*, *démarche*, *mdph*…).
La classe 6 pourrait être résumée par la phrase : « la *structure* va *mettre* en *place* des *mesures* d'*accompagnement* et de *continuité »*. La classe 5 est voisine et traite des pratiques confrontées à la crise (*stratégie*, *pratique*, *gestion*, *confronter*, *changement*, *inédit*…).
Enfin l'aide sociale à l'enfance est le cœur de la classe 12 : *enfant*, *famille* [d'] *accueil*, *parent, _ase* (aide sociale à l'enfance) … Si le discours de cette classe est surreprésenté chez une dizaine de témoins, l'un d'eux, un psychologue, le surutilise particulièrement puisque son discours est constitué pour moitié de segments de cette classe.
Les résidents qui ont témoigné partagent certains discours avec le personnel des foyers d'hébergement. Ainsi la classe 9 des moments conviviaux (*café*, *cour*…) est surreprésentée dans les deux sous-populations, montrant leur importance commune. Certains thèmes sont par contre légèrement surreprésentés chez les résidents (par rapport au personnel de foyer d'hébergement), comme celui des émotions (classe 10), sans atteindre l'importance de celles-ci dans le discours d'un agent de service hospitalier travaillant dans un Ehpad (donc également un

établissement accueillant des personnes sur le long terme) qui rend compte des sentiments des résidents de son établissement. Également surreprésenté chez les résidents est le discours de la classe des vacances (classe 3), le départ étant parfois vécu comme une libération : « je suis *allé prendre* le billet de train pour *aller* à _lieu_E. C'était *lourd* quand même d'être tout le *temps* chez moi. Là, pouvoir *prendre* mon masque et me dire je me *casse* c'est bien, je me suis *senti léger,* libéré » (Auteur 72). Par contre la classe 14 des gestes et des produits est surtout imputable au personnel, sans être exclusif, une résidente insistant sur les précautions qu'elle prend.

4. Évolution du discours sur la peur et le bonheur

Tous les scripteurs n'ayant pas commencé leurs écrits à la même période, l'analyse de spécificités d'une variable définie par les dates de confinement et de déconfinement ne rendrait pas compte de l'évolution du discours sur la peur et le bonheur. Nous avons, pour répondre à la question « y-a-t-il un noircissement du discours au fur et à mesure de l'écriture du journal ? » défini 4 périodes basées sur les quartiles de date de chaque témoignage, ne retenant que les témoignages ayant suffisamment de dates.

Le Tgen Peur est cité 461 fois dans le corpus, soit en moyenne une fois toutes les 2 pages. En moyenne un témoignage le cite 6.3 fois, mais sept témoignages ne le citent jamais (ce sont des témoignages relativement courts), et un le cite 38 fois.

Le Tgen Peur est très spécifique du premier intervalle interquartile (spécificité supérieure à 6), pour céder progressivement la place au Tgen BonheurMalheur au cours des 2 périodes suivantes, ce qui pouvait être attendu : l'inconnu inquiète (les lemmes *inquiet*, *inquiétude* et *panique* sont spécifiques de cette période), ensuite le poids des contraintes s'impose, provoquant un certain mal-être.

Avant de conclure, il convient de vérifier que cette évolution est largement partagée et ne dépend pas que d'un seul témoignage. La spécificité pour chaque témoignage du Tgen Peur au sein de chacune des 4 périodes a été calculée. Près de 60% des témoignages ont une spécificité positive durant la 1ère période, et seulement 25% durant la 2ème période, ce qui est une première confirmation. Nous avons ensuite défini, grâce à une classification ascendante hiérarchique, 5 profils d'évolution de la peur. Le premier concerne 27 témoignages, dans lesquels le discours sur la peur est toujours peu présent, le deuxième rassemble 14 témoignages ayant un profil moyen (donc comme celui tous témoignages confondus, confirmant la validité de l'évolution globale), le troisième 4 témoignages avec de fortes spécificités pour les 2ème et 3ème périodes, le quatrième 1 seul témoignage avec un profil atypique (très forte surreprésentation de la peur durant la dernière période), le cinquième avec 2 témoignages avec une très forte surreprésentation durant la première période. Lorsqu'on analyse l'avant-dernier

profil, on s'aperçoit que la dernière période est un récit de souvenirs : l'Auteur 58, à la fin de son témoignage, parle des ressentis du début. Il faut donc se garder d'assimiler jour d'écriture à jour narré.
Cette confirmation de la très forte présence de la Peur en début de journal ne doit pas faire oublier que la peur, généralement, est présente, quoique moins intensément, aux autres périodes.

5. Conclusions et perspectives

15 types de discours, répartis en 3 grand groupes, ont été mis en lumière par l'analyse textométrique. Ces discours sont parfois centrés sur un certain type de public (enfants, victimes d'addictions, proches du scripteur…), d'autre fois ou simultanément sur des ressentis plus généraux. Une analyse de l'évolution d'un ressenti, la peur, a complété cette description du corpus et il a été montré la nécessité de considérer certaines évolutions apparentes avec prudence, car le temps d'écriture ne correspond pas toujours au temps du ressenti.
Les deux autres ressentis, le bonheur/malheur et le dépit, n'ont été qu'effleurés et feront l'objet d'analyses futures.

Références

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Ratinaud P. (2009). Iramuteq: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires.URL: http://www.iramuteq.org.
Reinert M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, VIII(2) :187-198

Exploratory textual network for Italian Third Mission

Emma Zavarrone, Alessia Forciniti, Martha Friel

Università IULM – emma.zavarrone@iulm.it alessia.forciniti@iulm.it martha.friel@iulm.it

Abstract
In 2012, the Third Mission (*TM*) was introduced in the Italian university system such as institutional mission alongside the traditional missions of teaching and research. During 2013, the Minister of Education identified and proposed indicators and parameters aimed at periodic evaluation of the *TM*. In spite of the apparent simplicity of the *TM* definition, its declination and evaluation still raise many questions. The *TM* evaluation framework saw a series of trials which lead to a wide standardisation in the philosophy, methodologies and indicators, by defining an inappropriate model to establish the systematic comparisons between universities. The aim of this paper is to analyse the *TM* with a focus on the role of the universities to enhance the knowledge transfer in the socio-cultural and economic territorial environment. More precisely, by an exploratory textual network approach, we analysed the universities' strategic plans drafted from 2015 to 2022 for discovering how the *TM* has been depicted and incorporated in universities culture. We studied the importance attributed to the term "impact" in the definition of the strategic lines aimed at defining the *TM*. Our innovation in this contribution does not reside in the integration of textual analysis and Social Network Analysis but, in its conjoint application to the context of the *TM*, such as a constantly evolving socio-cultural and regulatory domain. The results of exploratory analysis suggest that currently, there is a mismatch between the evolving theoretical framework and the practice of academic activities drafted in the strategic plans. Thus, it results necessary the definition of other criteria to develop or improve the importance of the socio-economic repercussions deriving from the strategic lines in the Italian universities.

Keywords: third mission, strategic plans, textual network analysis, centrality measures

1. Introduction and background

The Third Mission (*TM*), as defined by ANVUR (National Agency for the

[1] VQR: Valutazione Quadriennale della Ricerca

Evaluation of the University and Research System in Italy), is an institutional mission of the university, alongside the traditional missions of teaching and research. The main aspects of Third Mission can be represented by: a) the cultural and social mission, through the increase of the well-being of society, in the educational, social and cultural fields; b) the economic enhancement of research, encouraging comparison and mutual development between the universities and stakeholders. The *TM* has been recognized as an institutional mission of the universities only in recent times and with a still incomplete normative forecast. The Legislative Decree 19/2012 constitutes the first organic reference that allows you to fully insert the *TM* in universities activities, while the Decree of the Minister of Education 47/2013 identifies the indicators and parameters of periodic evaluation of the *TM*. The uncertainty of the *TM* evaluation framework sees as the first trial the evaluation of research quality (*VQR*) from 2004 to 2010 by requiring universities to produce some indicators. Among the criteria, there are the activities that denote the inclination of the university to open up to the socio-economic context and to the enhancement and transfer of knowledge. That is to say, scientific and cultural dissemination activities, indicators of turnover on behalf of third parties, patents and spin-offs, and other extra activities related to research (training, dissemination, organization and management of courses or events). The experience of the VQR[1] has been useful to start the process but not decisive from the point of view of evaluation philosophy, methodology, indicators. The degree of standardisation of indicators has proved insufficient to establish systematic comparisons between structures. In 2015, ANVUR collected indicators at universities to build a comparable database, and on the basis of this system, an annual or biennial evaluation process has been initiated. At present, *TM* evaluation contributes to this process of periodic evaluation and is a self-assessment tool for universities and departments, as well as information and transparency on the many activities that universities carry out for the benefit of society. However, the quantitative indicators are not sufficient to describe the complexity of knowledge transfer and their relevant impacts on society. The adoption of academic activities must be interpreted by setting the outcomes in the context, such as demonstrated by Agasisti *et al.* (2019) measuring the spin–offs on local economic development. In particular, the input actions should take into account long-term economic and social returns. Understanding and measuring impact is gaining more importance in *TM* developments and to monitor the results of TM, universities should track longer term impact of academic activities, by contributing to the positive internal strategies developments in accordance to external benefits created. Thus, it is assumed that the Italian universities have taken on board the importance of the social and economic repercussions deriving from the strategic lines adopted. To analyze whether the concept of "impact" has become central in the definition of the *TM*,

we studied the strategic plans of the Italian universities by an integrated methodological approach which combines the textual analysis and Social Network Analysis in terms of visualisation and metrics.

2. Data, Methodology and Results

Data collection consists of the set of strategic plans drawn up by Italian universities from 2015 onwards, divided by size according to the Censis' ranking 2021 - 2022 of state (as mega, large, medium, small and polytechnics) and non-state universities (as large, medium and small). More precisely, for this exploratory study, we have analysed all strategic plans belonging to the classification of large state and non-state universities which have drafted the strategic lines by 2022. The methodology is oriented to an exploratory textual network analysis. Our innovative contribution does not lie in the integration of textual analysis and Social Network Analysis (SNA) (Wasserman & Faust, 1994) widely used in the literature, but in the intention to develop three different purposes: A) To combine the textual and linguistic aspect to study the semantics of the texts and to go beyond the coding of the bag-of-words that is limited to quantify the most frequent terms outside the context. B) To integrate two methodological approaches to detect the semantic of the text and to obtain models of sense through graph representation and properties of Social Network Analysis. C) To apply this methodological combination to TM, as a constantly evolving socio-cultural and regulatory context.

Data pre-processing was performed on 20 texts parsed and tokenized by means of the steps of normalization, lemmatization, removal of punctuation, numbers, stop words, universities' denomination. We detected 154,968 tokens and 120,456 types. The type-token ratio (TTR) indicates that the vocabulary variety is equal to 77.73%. The corpus consists of an average

sentence length of 97.75 and an average word length of 41.52. From the pre-treated corpus, we detected the bigrams, that are sequences of two adjacent elements from tokens, and we built an adjacency table Atxt which determines the co-occurrences between the bigrams, keeping a mark of their semantic relations. The adjacency matrix may be interpreted as a network structure, where the bigrams represent the nodes and the co-occurrences the edges. In this perspective, we visualise the most 70 frequent bigrams through the SNA (Figure 1) and, establish a semantic connection of the strategic lines of last years, helping us also with main normalised centrality measures (Table 1): degree centrality, betweenness centrality and closeness centrality.

Figure 1. Textual Network of top 70 bigrams

The main results of the exploratory textual network analysis show *TM* as the most central node. It represents 4.71% of strategic plans of larger state and non-state universities in Italy, demonstrating that the *TM* is becoming more valuable in the strategies for enhancing teaching and research. The graphic importance of *TM* is also demonstrated by the centrality measures: in terms of total number of edges and thus, normalised *degree centrality* (0.348); how number of times which the node lies in the shortest path between other nodes, in terms of *betweenness centrality* (0.544); finally, as *closeness centrality* (0.092) to the other nodes. Through the *Figure 1*, we observe how the *TM* is relegated to internationalization both in research and development areas such as digitalisation of resources, infrastructures of ICT and human capital. In turn, ICT is related to the realization of value and indicators aimed at the transfer of knowledge, one of the pillars of

TM. However, the results show that the hypothesis that the Italian universities have transposed the normative framework of the *TM* impact is rejected, because the economic and social repercussions are not frequent or central in the strategic documents.

Table 1. Normalised centrality measures of the bigrams

Bigram	*Degree*	*Closeness*	*Betweenness*
terza_missione	0.348	0.092	0.544
punteggio_soddisfazione	0.275	0.087	0.078
azione_indicatore	0.246	0.078	0.116
indicatore_formula	0.232	0.083	0.033
raggiungere_attività	0.203	0.087	0.073
termine_soglia	0.203	0.087	0.077
indicatore_target	0.174	0.082	0.087
attività_termine	0.174	0.087	0.050
valore_iniziale	0.159	0.063	0.002
raggiungere_punteggio	0.159	0.083	0.024

The exploratory nature of the methodological approach made it possible to analyse the *TM* and to assess a lack of alignment between the evolving theoretical framework and the practice of academic activities. In the future, our work aims to analyze the strategic plans of state and non-state universities of all sizes and to compare documents after 2022, in order to check whether there has been any evolution in the most recent strategic lines on the topic of the *TM* impact.

References

Agasisti T., Barra C. and Zotti R. (2019). Research, knowledge transfer, and innovation: The effect of Italian universities' efficiency on local economic development 2006-2012. *Journal of Regional Science*, vol. (59)5:819-849.

Anvur. (2015). La valutazione della terza missione nelle Università e negli Enti di Ricerca. Manuale per la Valutazione. https://www.anvur.it/wp-content/uploads/2016/06/Manuale%20di%20valutazione%20TM~.pdf

Censis (2022). La classifica Censis delle Università italiane (Edizione 2021/2022). https://www.censis.it/formazione/la-classifica-censis-delle-università-italiane-edizione-20212022/gli-atenei-statali.

Wasserman S. and Faust K. (1994). *Social Network Analysis.* Cambridge University Press.

SustDict: a customized lexical-based dictionary for CSR

Emma Zavarrone, Alessia Forciniti, Marta Muscariello

IULM University – emma.zavarrone@iulm.it alessia.forciniti@iulm.it marta.muscariello@iulm.it

Abstract
In 2001, EU defined Corporate Social Responsibility (CSR) as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis. Being socially responsible means not only fulfilling legal expectations, but also going beyond compliance and investing more into human capital, the environment, and the relations with stakeholders". Following this definition, the CSR' pillars are represented by environmental, social, and economic sustainability, and must be communicated to the society through appropriate reports. Sentiment Analysis (SA) can capture the tone and opinion of reports by detecting polarity and identifying the class of emotion hidden in the documents. SA can be conducted at different levels: document, sentence and aspect. However, SA depends on the availability of dictionary. This dictionary has a great influence on results and the corresponding analyses. The construction of a universal sentiment dictionary that fits all domains is very complex because word sentiment depends on the domain it has been used for. The dictionaries of given domain cannot easily be applied to other domains because terms can have opposite sentiments when used in different situations. For this reason, we developed a customized lexical-based dictionary for Italian sustainability terms using methods for sentiment polarity categorization based on polarity features, such as adjectives, verbs, and nouns, by combining textual analysis techniques with the properties of social network analysis. The corpus is composed by environmental, social, and governance (ESG) disclosure strategies for Italian listed companies that closed the financial year at 31st December 2021.

Keywords: sentiment analysis, lexicon-based methods, customized-lexicon, polarity classification, performance measures

1. Introduction

The awareness of the paradigm of sustainability (Lukman & Glavic, 2007) has enriched and diversified many fields of study, including the corporate organizational culture (Aguinis & Glavas, 2012). According to the European Commission (2011), the role of business concerning the sustainable development is related to activities based on corporate social responsibility (CSR) and thus

aimed at reducing the impact of the negative effects of business (Carpenter & White, 2004). At the same time, the sustainable development represents a model of strategic management to gain competitive advantage bringing the CSR to the level of a system of "governance" of transactions and relations between the company and its stakeholders, maximizing long-term profits (Chandler & Werther, 2014). The widest literature is based on the relationships between CSR and corporate performance (McWilliams & Siegel, 2011), while less developed is the integrated view of corporate sustainability management and CSR (Aguinis & Glavas, 2012). In this perspective, we are not able to clearly determine a well-defined language used by companies to communicate the CSR to stakeholders. One of the methods used to study communication and obtain an inclusive community perspective on a specific area is Sentiment Analysis (SA). In order to know the opinion on stock markets (e.g. Arnold & Vrugt, 2008), economic systems policies, SA finds application in fields such as business, organization, government. SA is performed by means of machine learning or lexicon-based methods or the combination of these two in a hybrid approach. In this paper, our interest is on the lexicon-based (LB) approaches, which according to some scenarios (e.g. Vassallo *et al.,* 2020) are preferred to supervised classification methods, since they are more robust for the classification across domains and present a greater scalability. LB methods allow the interpretation of a single word on its context-dependency and are, essentially, related to polarity lexicons which attribute a valence of positive, negative or neutral sentiment to a text (e.g. -1, 0, 1). However, the classification of polarity may depend on the domain in which each word fits and the lexicons of reference cannot cover all the meanings of the words specific to a domain. The field of application of SA offers different lexicons, some of which are based on general knowledge of sentiment by including different areas such as sports, politics, language of social media (Nielsen, 2011). Other lexicons are domain-based in sector specificities such as the economic-financial environment (Henry, 2008; Loughran & Mcdonald, 2011). If on the one hand the lexicons of the sentiment of general knowledge have the advantage of a wider coverage of the phenomena, on the other hand, they have lower performance than the domain-based lexicons, since some terms may neither be present in a general lexicon nor can have a different contextual classification. Technically, customized lexicon requires a manual labelling data often very expensive. Thus, pre-constituted dictionaries specific to a language and a domain of interest are increasing, as well as widespread used. These dictionaries are useful in contexts where the available data are reduced. This perspective is very common for low-resource languages such as Italian. Different general lexicons of sentiment have been proposed for Italian, they are often the results of translation from English. Among the most popular, we can mention SentiWordNet (Esuli & Sebastiani, 2006) and the National Research Council

(NRC) Emotion lexicon (Mohammad & Turney, 2010). The same limitations are connected to domain-based lexicons mainly developed in English (cf. Henry, 2008; Loughran & Mcdonald, 2011) but, with reference to sustainability management there is a lack of resources both in English and other languages. Thus, our contribution aims to select an effective method for approaching SA on CSR in Italian. The paper consists in a practical evaluation of the performance of LB methods by means of two steps: the first step is oriented to assess the use of general lexicon, and the second one is focused on the evaluation of a customized-lexicon given by the implementation of general lexicons with lexical items connected to CSR. The lexical specificities are detected combining textual and Social Network Analyses and specifically we compare SentiWordNet and NRC Lexicon.

2. Method

The method under development goes into the direction of proposing an innovative model to improve the lexicon performance and implement the lexical items related to CSR in Italian corporate communication. Our innovative methodology is based on the building of a customized-lexicon by means of a multi-stage model that combines text analysis with Social Network Analysis (SNA). The main advantage of our model is the possibility to detect unigrams and bigrams which characterize a given corpus and to attribute them a positive or negative valence performing a syntactic and semantic analysis of each lexical item selected. For this purpose, we used a cross validation procedure based on the sub-setting the corpus in five sub-corpora, each one composed by a random sample of 26 CSR reports. We develop the model (Figure 1) in sample 1 (train set) and we use the rest of samples to test the method (test set). The stages of the method are: 1) corpus creation and a document-term matrix $\mathbf{DTM}_{dxt}$, with d CSR reports, and t terms, construction. The data pre-processing operations have been implemented through typical operations of removal of punctuation, numbers, stop words and company's name, and lemmatization. Specifically, we built the **DTM** normalizing the schemes for weighting from tf (term-frequency) to tf-IDF (term frequency-Inverse Document Frequency) so as to quantify the specificity of a term in inverse proportion to the number of documents in which it occurs; 2) Transformation **DTM** in an adjacency matrix, $\mathbf{M}_{txt}$, based on the study of co-occurrences which allows to keep mark of the structure of each text without breaking the links with semantic analysis. We take advantage of social network features to detect unigrams and bigrams using the centrality measure known as Freeman's closeness C_{ci}, (Freeman, 1978) which represents the shortest path from one to the other; 3) All terms greater to the median value of closeness represent the words to implement into lexicon; 4) We compare the C_{Ci} selected terms with lexicon terms and we implement a decision rule based on

the presence/absence paradigm. If the C_{Ci} terms already do exist then the procedure will be stopped, otherwise the return to the syntactic and semantic analysis that precedes the pre-treatment of the text will be realized; 5) By means of analysis both of collocations among the terms and the context (kwic - keyword in the context), we assign a sentiment label. The new sentiment (positive or negative) will be adjunct to the k'. The k' will be tested in four samples (i.e. testing set). We applied performance measures for binary sentiment classification to validate our method. More precisely, we estimated the overall effectiveness of the classifier in terms of accuracy; the effectiveness of the classifier to identify positive labels as sensitivity or recall; the effectiveness of the classifier to determine the negative labels in terms of specificity.

STEP	DESCRIPTION
1	$From\ DTM\ to\ M_{txt}$ $M_{txt}\ with\ \ \mu_{ij}\begin{cases}1 & if\ there\ is\ an\ edge\ from\ term\ i\ to\ term\ j \\ 0 & otherwise\end{cases}$
2	$Compute\ C_{c(i)}$
3	$Let\ X = C_{c(i)} > \text{Median}\ C_{c(i)}$
4	$if\ X\begin{cases} = Dict(k) & stop. \\ \neq Dict(k) & go\ to\ KWIC(X)\end{cases}$
5	$Compute\ X_1 = sentiment\ vector\ (X - X(Dict(k))$
6	$Add\ X_1\ \ to\ Dict(k)$

Figure 1. Algorithm structure: steps

3. Data and results

We collected the sustainability reports drafted by the complete listing of Italian companies listed on stock exchanges that closed the financial year at 31st December 2021, in accordance with the Legislative Decree No. 254 of 30th December 2016[1]. Of the complete list that includes 196 companies, we have selected 130 documents, because of the unavailability of some sources or the drafting of the report in English. At first, we create the corpus and applied the

[1] https://www.borsaitaliana.it/notizie/sotto-la-lente/informazioni-non-finanziarie.htm

data treatment procedure in order to create a vocabulary. After sub-setting the corpus, we analysed the vocabulary of each sample through an exploratory textual data analysis (ETDA (Table 1). We observed – for each sample – a generalist vocabulary with few references to a specialized language or related to technicalities of the domain. To apply our model and to verify the effectiveness of the customized-lexicon on CSR, we propose the comparison of performance of two of the most popular general lexicons (indicated by Dict(k)) for Italian before and after their implementation. The first lexicon used is SentiWordNet (k=1) developed by Esuli & Sebastiani (2006) from a corpus-based semantic approach to detect positive, negative and objective (neutral) polarity. The second lexicon used is NRC (k=2) Emotion Lexicon developed by Mohammad and Turney in 2010. NRC lexicon is used to detect simultaneously the semantic orientation (positive or negative) and the emotions. Mohammad and Turney used the Mechanical Turk (Amazon Service) obtaining human annotations at sense level rather than at word level in an efficient way. In SentiWordNet scores are measured on a continuous scale with different intensity from 0 to 1. NRC Emotion Lexicon presents instead a classification based on binary polarity negative or positive (-1 or 1). Thus, to perform the comparison between two lexicons, we converted each score of SentiWordNet into a binary sentiment category labelled as negative or positive, by using only the sign of the score and removing the polarity equal to 0 marked as objective or neutral.

Table 1. ETDA

Sample	*N. lemmas*	*Unigrams*	*Bigrams*
First	21,754	gestione, gri, rischio, emissione, …	consumo_energetico, capitale_umano, codice_etico, lotta_corruzione, …
Second	21,867	salute, consumo, impianto, sicurezza, …	aspetto_materiale, gestione_rischio, pari_opportunità,corporate_governance, …
Third	22,496	modello, capitale, formazione, obiettivo, …	controllo_rischio, collegio_sindacale, risorsa_umana, ...
Fourth	20,915	mercato, salute, impatto, gri, …	gestione_responsabile, ricerca_sviluppo, fonte_rinnovabile, donna_uomo, …
Fifth	20,644	ambientale, sviluppo, processo, sistema, …	salute_sicurezza, controllo_interno, materia_prima, etica_integrità, …

The overall sentiment of 130 sustainability reports showed a predominance of positive sentiment both by means of SentiWordNet and NRC Emotion Lexicon.

More precisely, we detected 257,207 negative words and 508,598 positive ones by using the SentiWordNet lexicon; while we observed 120,905 negative words and 397,615 positive ones with NRC. We obtained the same prevalence of positive sentiment for each sample, varying equally between the two lexicons from 19% to 22%. The negative polarity is also almost distributed in a similar way between the two lexicons varying from 18% to 23%.

Table 2. Performance measures of the samples before the implementation

Sample	*Accuracy*	*Recall*	*Specificity*	*Accuracy*	*Recall*	*Specificity*
First	0.53	0.45	0.69	0.68	0.30	0.78
Second	0.53	0.38	0.69	0.64	0.28	0.77
Third	0.51	0.37	0.68	0.53	0.28	0.80
Fourth	0.55	0.36	0.69	0.66	0.27	0.79
Fifth	0.65	0.36	0.70	0.52	0.26	0.79

The performance measures computed on both lexicons before the implementation (*Table 2*) show that NRC lexicon detected on average a better accuracy in the sentiment classification. Its average accuracy is greater of 5.2% than SentiWordNet's performance. This depends on its specificity, that is NRC's ability to better detect the words with negative polarity in this context for 78.6%. On the contrary, SentiWordNet well performed on the positive side, showing a greater average percentage of recall equal to 38.4%. Our model allowed to detect 65 unigrams and 82 bigrams greater to the median of closeness. By the annotation procedure aimed to attribute positive or negative valence to each item, we determined 57 positive unigrams and 8 negative ones, 61 positive bigrams and 21 negative ones. The implementation of the lexicons through the new 147 lexical polarized items has determined that the customized-lexicon of SentiWordNet is composed of 20,240 items and NRC Lexicon consists of 5,615 items. The performance measures on the implemented lexicons showed an overall improvement (*Table 3*) which on average varies between the different measures from 0.8% to 8.8%. The *Table* 3 confirmed a 2% higher mean accuracy for NRC lexicon. However, despite the greater effectiveness of the NRC classifier in identifying negative labels, the implementation of the lexicon has mostly improved the SentiwordNet's specificity by 0.6% more than the NRC. The domain-based implementation favoured the performance of SentiWordNet for 5.2% compared to 4.6% of NRC. The performance in the classification of positive labels has instead confirmed the greater effectiveness of SentiWordNet as recall, since NRC increased on average its recall by only 0.8% compared to 2.8% of SentiWordNet.

Table 3. Performance measures of the samples after the implementation

Sample	*Accuracy*	*Recall*	*Specificity*	*Accuracy*	*Recall*	*Specificity*
First (baseline)	0.59	0.52	0.76	0.69	0.32	0.84
Second	0.60	0.39	0.74	0.71	0.28	0.83
Third	0.57	0.39	0.69	0.70	0.28	0.85
Fourth	0.68	0.39	0.77	0.68	0.28	0.81
Fifth	0.68	0.39	0.75	0.69	0.27	0.83

4. Conclusions and future work

We studied the paradigm of sustainability in the Italian corporate culture, by analysing the sentiment of the sustainability reports drafted by listed companies. We proposed an innovative method for developing a customized lexical-based dictionary able to investigate the sustainability in Italian language. By integrating bag of words coding, linguistic approach and SNA, we implemented and compared two Italian general lexicons. To test the efficacy of our model, we split under cross validation perspective our corpus in five random samples: one used such as train set and four as test set. The procedure highlights a significant improvement of performance measurements in all samples compared with the baseline sample. Future developments lead us to include other sources to improve data training and apply a neural network approach. The development of SustDict package, based on the algorithm structure proposed, represents a further step to deal with.

References

Aguinis H, Glavas A. (2012). What we know and don't know about corporate social responsibility: A review and research agenda. *Journal of Management,* vol. (38): 932-968.

Arnold I., Vrugt E. (2008). Fundamental uncertainty and stock market volatility. *Applied Financial Economics*, vol. (18): 1425-1440.

Carpenter G. and White, P. (2004). Sustainable development: Finding the real business case. *Corporate Environmental Strategy. International Journal for Sustainable Business*, vol. (11): 51-56.

Chandler D. and Werther W.B. (2014). *Strategic Corporate Social Responsibility. Stakeholders, Globalization, and Sustainable Value Creation*. 3ª ed. SAGE Publications.

Esuli A. and Sebastiani F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *In Proceedings of LREC 2006.*

European Commission (2011). A renewed EU strategy 2011-14 for Corporate Social Responsibility.

Freeman L.C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, vol. (1), pp.215-239.

Henry E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written?. *Journal of Business Communication* (45): 363-407.
Loughran T. and Mcdonald B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance (66.1): pp. 35-65.
Lukman R. and Glavic P. (2007). Review of sustainability terms and their definitions. *Journal of Cleaner Production*, vol. (15): 1875-1885.
McWilliams A. and Siegel D.S. (2011). Creating and capturing value: Strategic corporate social responsibilty, resource-based theory, and sustainable competitive advantage. *Journal of Management*, vol. (37): 1480-1495.
Mohammad S. and Turney P. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*
Nielsen F.A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *In Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages*, pp. 93-98.
Vassallo M., et al. (2020). Polarity Imbalance in Lexicon-based Sentiment Analysis. *In Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*. Accademia University Press.

Construction of an encyclopedic dictionary of tourism from scientific corpus

Lichao Zhu[1], Fabrice Issac[2],
Touria Ait El Mekki[3], Olivier Hû[3]

[1]Université Paris Cité – lichao.zhu@gmail.com
[2]Université Sorbonne Paris Nord – fabrice.issac@gmail.com
[3]Université d'Angers – touria.aitelmekki@univ-angers.fr
olivier.hu@univ-angers.fr

Abstract
In this article, we present our method in order to create a nomenclature while respecting the choice of the experts, not only to identify the terms but also to be able to account for their origins and their specificities. Starting by compiling a corpus of scientific articles from various electronic tourism journals, we labelled the corpus with certain compound words and proper nouns were identified in it. We then implemented different strategies to identify candidates using syntactic patterns. A particular attention was paid to the presentation of syntactic and statistical results for experts in the field of tourism (interface and terminology sheet).

Keywords: textometry, tourism, scientific corpus, linguistic modeling, supervised learning.

Résumé
Nous nous présentons dans cet article notre méthodologie pour la création de la nomenclature en respectant la volonté des experts, de non seulement identifier les termes mais aussi d'être capable de rendre compte de leurs origines et de leurs spécificités. Nous avons commencé par constituer un corpus d'articles scientifiques issus de différentes revues électroniques du tourisme. Ce corpus a été étiqueté, certains mots composés et noms propres y ont été identifiés. Nous avons ensuite mis en œuvre différentes stratégies pour identifier des candidats en utilisant des patrons syntaxiques. Nous avons porté une attention toute particulière à la présentation des résultats syntactico-statistiques pour les experts du domaine du tourisme (interface et fiche terminologique).

Mots clés: textométrie, tourisme, corpus scientifique, modélisation linguistique, apprentissage supervisé

1. Introduction

La recherche se définit en discipline mais aussi, et peut être de plus en plus, en domaine ou en champ. La lexicométrie, le TAL ou les humanités numériques

sont des exemples de ces domaines qui mettent en œuvre plusieurs disciplines. Se crée alors un « langage commun » formé de termes issus des différentes disciplines et de termes forgés pour le domaine ou le champ lui-même. Ces termes sont issus d'un consensus : ils sont le produit d'une littérature scientifique et d'un temps de maturation importants. Le domaine du tourisme en France est une évidence économique et sociétale. Il est également un champ de recherche transdisciplinaire (économie, géographie, histoire, sociologie…), mais qui ne disposait pas ou peu de structures dédiées à la recherche. Démarré en 2019, le Groupement d'intérêt scientifique (GIS) *Études touristiques* entend participer à combler ce manque et propose de regrouper les chercheurs pour lesquels le tourisme est la thématique centrale. Dans cette optique la direction du GIS a initié un travail épistémologique avec, entre autres, l'élaboration d'un ouvrage de référence de type dictionnairique et/ou encyclopédique. Pour répondre à cette problématique nous avons constitué un corpus d'articles scientifiques issus de différentes revues en ligne du tourisme. Ce corpus a ensuite été étiqueté, certains mots composés et noms propres y ont été identifiés. Cette ressource nous permet de mettre en œuvre dans un premier temps différentes stratégies classiques pour identifier des candidats (tf-idf, fréquence, spécificité, information mutuelle…) en utilisant des patrons syntaxiques. Nous suivons en cela une démarche proche de celle proposée avec l'outil *Termostat* (Drouin 2003).

2. Contexte

Dans le cadre de la création de cette encyclopédie du Tourisme, la direction du GIS a tout d'abord structuré sa démarche avec la création d'un comité éditorial en parallèle du comité scientifique déjà existant. Ce comité éditorial est composé de différents experts du GIS, volontaires et co-optés dans un souci de pluralité des disciplines représentées sans pour autant prétendre à l'exhaustivité. Ce comité éditorial a en charge : 1. l'appel à participation auprès des chercheurs du réseau du GIS afin que les auteurs potentiels proposent des termes-candidats ; 2. la sélection des entrées du dictionnaire encyclopédique parmi les termes-candidats proposés par les membres du GIS ; 3. la définition des règles éditoriales auxquelles les auteurs devront se soumettre : taille des articles, usages des illustrations, gestion des citations… ; 4. le suivi éditorial et les relations avec les auteurs : relances, modifications, validations. Pour éviter la posture d'expert et le piège d'une analyse critique restreinte aux disciplines des membres du comité éditorial, la direction du GIS a décidé d'initier un travail d'analyse de la production scientifique du domaine. Il a donc été entamé la construction d'un corpus représentatif de la production scientifique en Tourisme dont l'exploitation doit répondre à deux objectifs : extraire automatiquement des unités lexicales qui pourront être proposées au comité éditorial ; enrichir la compréhension et l'appréhension par le comité éditorial d'un candidat-terme via la mise en

évidence de son usage au sein du corpus. La volonté des experts est de non seulement identifier des termes mais aussi d'être capable de rendre compte de leurs origines et de leurs spécificités notamment disciplinaires. Pour ce travail, nous nous intéressons donc à l'extraction de termes spécialisés en tourisme issus d'articles scientifiques. L'idée est de proposer des termes à valeur thématique (*thematic valuable words*) à partir de méthodes qui allient des approches statistiques et modélisation lexicographique et syntaxique. Nous utilisons ensuite les annotations du comité éditorial pour améliorer, dans une approche supervisée, la précision des propositions de termes. Par rapport au schéma de traitement classique, nous ne recourons pas à une référence déjà existante (Nazarenko et al. 2009) mais nous nous appuyons sur notre corpus et les avis de nos experts.

2.1. Lexicographie et dictionnairique

Même si nous considérons que le travail réalisé ici comporte un volet encyclopédique, les outils et méthodologies à mettre en œuvre se rapportent principalement à la lexicographie. Le volet dictionnairique (au sens de Pruvost 2005) ne doit cependant pas être oublié et sera traité dans un second temps. Les différents aspects à prendre en compte pour la réalisation concrète du dictionnaire encyclopédique, notamment sur la question de l'usage ne doivent cependant pas être complètement occultés. Nous nous concentrons ici sur deux tâches : construire une nomenclature et proposer aux experts des outils permettant d'en établir leur sens. Les apports de l'informatique à la lexicographie et à la dictionnairique sont bien connus. On trouvera par exemple dans Béjoint 2007, Jacquet-Pfau 2005 et Bertin 2009 un certain nombre de points militant pour l'utilisation à tous les niveaux de l'outil informatique (construction, représentation, exploitation). Parmi tous les points abordés, le corpus tient une place centrale. Concernant son utilisation retenons qu'il permet : d'établir une nomenclature représentative ; d'avoir une information fidèle à l'usage ; de travailler en synchronie mais aussi en diachronie ; et d'exhiber des unités lexicographiques complexes.

3. Nomenclature

L'élaboration de la nomenclature est le premier pas vers la construction des dictionnaires. Il s'agit de définir le lexique sinon exhaustif du domaine du moins représentatif. Nous ne nous trouvons pas ici dans le cas d'un dictionnaire général et la tâche s'en trouve simplifiée. Cependant cette notion de représentativité n'en reste pas moins posée et c'est le corpus de par son volume et son origine qui dans un premier temps sera utilisé pour construire la nomenclature. Cette étape pose un grand nombre de questions pour tous les types de dictionnaires. Il s'agit finalement de définir quelles informations vont être présentées. C'est une tâche importante au cours de laquelle le lexicographe doit se poser un grand nombre

de questions (De Villers 2011). Il faut non seulement établir une liste de candidats mais aussi : (1) Placer les termes sur un axe spatial (y a-t-il des variations sémantiques selon la région) et/ou temporel (le sens a-t-il changé au cours du temps). (2) Garder/rejeter le candidat selon son niveau de spécialisation (L'homme and Polguère 2008). (3) Décider de critères d'insertion des emprunts. Ce questionnement présuppose la construction d'une première liste de candidats. Celle-ci peut être construire en (i) faisant appel à l'introspection des experts (ii) en utilisant une nomenclature existante (iii) en utilisant des outils statistiques permettant d'extraire les termes saillants. D'un point de vue statistique la fréquence et la répartition ne peuvent être les seuls critères. Les experts le sont dans leur domaine, dans le champ du tourisme, mais ne sont pas des lexicographes. A nous de proposer des candidats, d'expliquer pourquoi statistiquement ils ont une importance et de laisser au groupe d'experts le choix sur : Le terme monolexical ou polylexical, ses variantes flexionnelles, pourquoi il a été sélectionné (quelle mesure) et sa répartition au sein du corpus de manière absolue et en diachronie.

4. Élaboration d'un corpus d'articles scientifiques

Le corpus est issu de trois revues scientifiques[1], identifiées par le comité scientifique du GIS comme représentatifs du champ, de par leur large couverture pluridisciplinaire et leur ancienneté. La récupération s'est faite par aspiration des pages du site, ce qui a permis d'ajouter tout un ensemble de méta données (auteur, le cas échéant traducteur, bibliographie…). L'ensemble du corpus est structuré en XML en utilisant une DTD spécifique. Un tel corpus est a priori représentatif de la recherche en tourisme pas du tourisme et de ses acteurs. Le vocabulaire et les usages professionnels en sont par définition exclus. L'ensemble de ces informations permet d'envisager un traitement des éléments lexicaux sous différents axes. Il est ainsi possible, et pertinent pour notre étude, d'observer l'évolution du lexique en diachronie. Les textes d'article extraits sont tokénisés et étiquetés à l'aide de l'outil *Corpindex*[2]. Cet étiquetage se fait par application de dictionnaires généraux et une levée d'ambiguïté à partir de règles ce qui rend le corpus exploitable au niveau des parties du discours. Au total nous disposons d'une ressource composée d'articles réparties sur 20 ans comportant plus de 4 millions de mots. Si ce corpus n'est pas la ressource *définitive* par rapport au domaine du tourisme nous estimons qu'il le représente de manière satisfaisante mais c'est encore une fois les experts qui nous diront s'il faut en reconsidérer les contours.

[1] Mondes du tourisme : https://journals.openedition.org/tourisme/, Teoros : https://journals.openedition.org/teoros/ et ViaTourism : https://journals.openedition.org/viatourisme/
[2] https://github.com/ProfessorFabrice/Corpindex

5. Méthodologie

Nous indiquons ici les différentes mesures, ou stratégies, mises en œuvre afin de constituer des listes pertinentes de candidats. On distingue les traitements sur des unités monolexicales (p. ex. fréquence), polylexicales (p. ex. information mutuelle) et transversaux (répartition). L'ensemble des traitements ne présente pas de caractéristiques particulières et est classique dans le cadre d'outils terminologiques. Notre objectif est de construire une fiche descriptive, la plus complète possible, qui sera soumise aux experts.

5.1. Mesure de dispersion

Le corpus est trié chronologiquement. De ce fait, une recherche de l'ensemble des occurrences d'un terme permet à la fois de mesurer sa répartition (cette mesure est calculée à l'aide d'un calcul du coefficient de variation obtenu par le rapport entre l'écart-type et la moyenne) et son évolution dans le temps. Il est à noté que cette mesure considère le corpus comme étant monolithique. Plus cette mesure est faible plus le terme est présent sur l'ensemble du corpus. À titre d'illustration, le mot de, qui est extrêmement bien réparti, a une mesure de 0,03, le terme *chaîne hôtelière* à une mesure égale à 1,8 qui indique une répartition relativement uniforme à l'inverse de touriste postmoderne avec une mesure de 3,47. Nous nous sommes intéressé aux scores TF-IDF (Salton and Buckley, 1988) des noms dans l'ensemble des documents. Ainsi, nous avons retenu le meilleur score (t_k) et la moyenne des scores (t_k) pour chaque nom monolexical. Ces mesures que nous proposons aux experts aident à faire émerger des noms (unités monolexicales) à l'aune des fréquences en mesurant l'importance d'une occurrence dans un document et son importance dans l'ensemble des documents. Le croisement des deux suppute qu'un mot a de la valeur thématique. Ainsi, parmi les 20 noms qui ont obtenu les meilleurs scores, nous remarquons des mots tels que *surf*, *nautism*, *weather*, *tabasco* et *hospitalité*. Il est à noter que cette mesure est à prendre avec précaution. Comme l'indique Claveau (2012) il existe des mesures, comme Okapi, qui semblent plus efficaces. Notre indicateur est donc une baseline qu'il faudra ajuster.

5.2. Unités polylexicales

La recherche de collocations est motivée par le fait qu'elles sont un bon indicateur du genre des textes (Spina and Tanganelli 2012), la notion de genre étant ici associée aux articles scientifiques liés au champ du tourisme. Par ailleurs, ces collocations permettront d'identifier des unités monolexicales *productive* et par conséquent potentiellement intéressante pour la nomenclature. Nous utilisons, arbitrairement, deux mesures afin de proposer aux experts un ensemble d'unités polylexicales : le z-test et l'information mutuelle. Nous envisageons, si cela s'avère nécessaire, d'utiliser d'autres mesures comme la mesure de gravité

lexicale (Daudaravičius and Marcinkevičienė 2004). L'intérêt de disposer de deux mesures est le fait que les résultats obtenus présentent un certain nombre de différences (sur les 200 premières unités trouvées par les deux méthodes, moins d'un quart sont identiques). Le z-test propose une mesure capable de rendre compte du fait que les mots qui composent une collocation apparaissent plus fréquemment ensemble que ne le voudrait le hasard. La méthode probabiliste basée sur l'information mutuelle, quant à elle, calcule la quantité d'informations partagée par deux mots. Dans les deux cas nous filtrons a priori les couples à partir de patrons syntaxiques. L'utilisation des résultats permet en outre d'identifier soit des prédicats soit des arguments pour des entrées données :

- *(pratiques, opérateur, aménagement, activité, expériences) touristiques*
- *espace (naturel, public, naturels, publics, rural, urbain)*

6. Présentation des livrables

A l'issue des différents traitements réalisés nous proposons deux types de résultats aux experts. Le premier est une fiche synthétique construite pour certaines unités (voir figure 1). Le deuxième est une interface de type concordancier qui en outre permet de retourner facilement au texte complet (voir figure 2). Ces deux ressources doivent à la fois refléter avec exactitude la réalité du corpus et être accessible pour un public non familiarisé avec ce type de résultats.

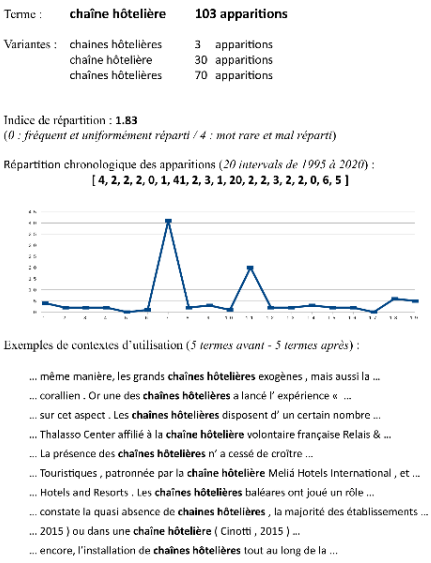
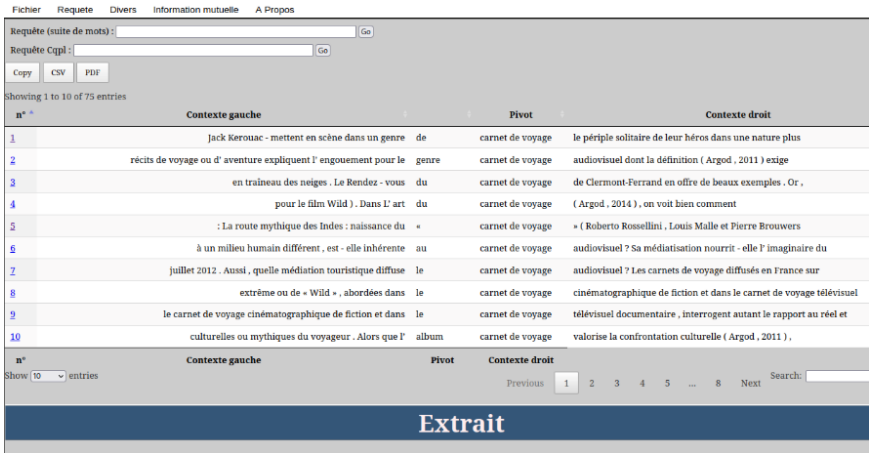


Figure 1. Exemple de fiche synthétique d'information

Figure 2. Capture d'écran de l'outil d'exploration du corpus



7. Conclusion, discussion, perspectives

Le travail présenté ici est le premier volet d'un projet visant à co-construire un dictionnaire encyclopédique capable de couvrir le champ du tourisme : la nomenclature. C'est une problématique qui nécessite évidemment les compétences d'experts du domaine mais également en lexicographie et, pour ce volet particulier du projet, de TAL. La méthodologie mise en œuvre consiste à développer des outils heuristiques de découverte de candidats au sein d'un corpus constitué d'articles scientifiques. Dans un deuxième temps les candidats sont présentés aux experts (fiche descriptive, outil d'exploration du corpus) et les raisons de l'acceptation ou du refus sont explicités. Par exemple les mots *voyage* et *station* malgré des statistiques les classant parmi les termes les plus pertinents, n'intéressent pas les experts qui les jugent trop imprécis. Ils préfèrent les termes *station verte* ou *station intégré*e malgré un classement bien plus faible. De la même manière les termes *station balnéaire* également bien placés ne seront pas choisis car considérés comme une instance de *station intégrée*. Une fois les termes sélectionnés la démarche est donc double. D'une part un processus itératif pour améliorer nos outils d'extraction. Comme nous l'avons évoqué nous confronterons les termes ainsi triés et ceux annotés et retenus par les experts afin de proposer une stratégie d'apprentissage supervisé intégrée dans notre processus de découverte de nouveaux candidats.

D'autre part, un travail de formation et d'information en direction des experts dans le cadre de leur travail épistémologique sera nécessaire. Ils vont devoir formaliser leur démarche, accompagné en cela par les outils que nous mettons à leur disposition afin de les aider à appréhender les usages terminologiques de la communauté des chercheurs en Tourisme.

Références

Henri B. (2007). Informatique et lexicographie de corpus : les nouveaux dictionnaires. *Revue française de linguistique appliquée*, 12(1), pp. 7-23.

Annie B. (2009). Grammaire et dictionnaire : le parti pris des mots. *Linx. Revue des linguistes de l'université Paris X Nanterre*, 61, pp. 71-85.

Vincent C. (2012). Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf (vectorization, okapi and computing similarity for nlp : Say goodbye to tf-idf) [in french]. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, *volume 2 : TALN*, pp. 85-98.

Vidas D. et Ruῑta M. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2), pp. 321-348.

Marie-Éva D. V. (2011). *Profession lexicographe*. Les Presses de l'Université de Montréal.

Patrick Drouin. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1), pp. 99-115,

Christine J.P. (2005). Pour un nouveau dictionnaire informatisé. *Ela. Études de linguistique appliquée*, 1, pp. 51-71.

Marie-Claude L. et Alain P. (2008). Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*, pp. 191-206.

Adeline N., Haifa Z, Olivier H. and Jonathan V. P. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Revue TAL, 50, varia*, pp. 257–281, URL https://hal.archives-ouvertes.fr/hal-00516698.

Jean P. (2005). Quelques concepts lexicographiques opératoires à promouvoir au seuil du xxie siècle. *Éla. Études de linguistique appliquée*, 137(1), pp. 7-37.

Gerard S. et Christopher B. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5), pp. 513-523.

Stefania S. et Elena T. (2012). Collocations as an index for distinguishing text genres. *Corpus*, 11, 01.