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 1
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 979-12-80153-30-2

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 1

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*The Italian Social Mood on Economy Index: recent methodological developments* **213**
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.

BIBLIOMETRIC AND SCIENCE MAPPING ANALYSIS: MODELS, SOFTWARE TOOLS AND FUTURE CHALLENGES

Manuel J. Cobo

Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada – mjcobo@decsai.ugr.es

Abstract
Nowadays, the scientific community generates an immense amount of information in the form of articles or other forms of knowledge. Processing this enormous amount of information to extract the underlying knowledge is an arduous task, which requires the use of appropriate tools. In this sense, and from a traditional point of view, bibliometrics has been in charge of the study of scientific publications, both in terms of production and impact. On the other hand, the analysis of scientific maps is responsible for modeling the scientific system as a network, in order to subsequently analyze and understand it as a complex system. Thus, in this contribution we will describe the main techniques, models and software tools for the analysis of large volumes of information. In addition, we highlight some future lines of work that will need to be addressed in the coming years.

Keywords: bibliometrics, scientometrics, science of science, science mapping analysis

1. Introduction

Nowadays, innovation, development, and research activities, and in particular science and technology, generate an immense number of documents and data in the form of publications, such as scientific articles, patents, theses, contributions to conferences, technical reports, or project reports, among others, forming a global, dynamic and complex knowledge structure. In addition, the total volume of scientific and technological information is increasing year by year, following a constant growth trend. From a traditional point of view, information is usually indexed in bibliographic databases (Web of Science, Scopus, etc.), patent databases (European Patent Office, United States Patent and Trademark Office, etc.),

or repositories (open access or private) such as arXiv. Thus, one of the possible ways of analyzing knowledge could be by measuring academic impact, i.e., it can be reduced to quantifying the number of documents that cite another in a given database. However, it is increasingly common for scientific and technical actors to disseminate their activity in academic (Mendeley, ResearchGate, etc.) or generalist (Twitter, Facebook) social networks, which produces a quantifiable social impact through the so-called altmetrics (Costas et al. 2015).

In this sense, the global science and technology system can be understood as a complex structure (Batagelj and Cerinšek 2013, Cobo et al. 2011a) in which academic and industrial actors, and the global knowledge generated, relate to each other, and can be represented as a network or graph, which could be called the Graph of Knowledge (GRAKO). Within this structure there are a multitude of complex relationships between the different types of nodes or objects of the GRAKO. Thus, there can be, among others, nodes related to people (researchers), institutions, countries, documents, concepts, journals, among which a wide variety of relationships can be established: social relationships based on collaboration, training, and membership (to specific groups), relationships based on similar concepts that form unique lines of research, relationships based on intellectual relationships (citation), etc. Thus, some of the relationships may be visible and/or simple, and others may be complex and/or not visible. Moreover, in GRAKO, relationships are heterogeneous and dynamic, i.e., relationships are not only established between two nodes of the same type, but various types of nodes may be involved, and these relationships may be changing over time. Thus, through the in-depth analysis of GRAKO, it is possible to understand what the research and development process is, what aspects make it successful (Sinatra et al. 2016), what similarities between conceptually distant research may occur, what have been the disruptive topics in a line of research, how knowledge is transferred to society, how society can help in research and development, how the socio-economic and geographic aspects of a region can affect research (Lebel and McLean 2018), how research can affect socio-economic indicators, etc. In this sense, the meta-science called *"Science of Science"* (Fortunato et al. 2018), allows us to define, model, analyze and interpret the GRAKO using tools from statistics, sociology, social sciences, computer science, artificial intelligence and complex systems.

The GRAKO could be analyzed, and therefore, understood, using different models, methodologies, algorithms, and software tools. In that sense, in this contribution, we review and summaries the main methods and software tools able to deal with scientific knowledge and the future trends and open problems that exist today.

2. Bibliometric analysis

Traditionally, the GRAKO has been studied from a bibliometric perspective (Cobo et al. 2011a), which employs mathematical models to determine the size, growth and distribution of scientific and technological literature. Bibliometrics contains a branch dedicated to the design of indicators to measure the scientific impact of the actors involved in innovation, development and research activities, since these allow an objective evaluation of the actors, provided that certain guidelines are met (Hicks et al. 2015).

Mainly, bibliometric could measure two wide areas. For the one hand, it is delved to productivity indicators, such as, the number of documents an author published, the percentage of documents in a given database, the number of documents in a specific author position, etc. On the other hand, visibility an impact indicators are based on the accountability of the citations received by the documents, journals, or authors, among others levels. It is important to clarify that impact indicators must be used at the correct level, and always put them into context. That is, the well-known JCR is a journal level indicator that should be used to measure the supposed quality of a journal, and not to measure the quality of a given paper. On the other hand, a single article could be compared with those in a similar category and in the same year in order to know if the amount of citations received make some different. For example, if the article is among the 1% most cited, it is classified and highly cited paper. Finally, at authors level we should highlight the h-index (Hirsch 2005) and its derivates (Alonso et al. 2009).

3. Science mapping analysis

Science mapping or bibliometric mapping is a spatial representation of how disciplines, fields, specialities, and documents or authors are related to one another (Cobo et al. 2011a). It is focused on monitoring a scientific field and delimiting research areas to determine its cognitive structure and its evolution. Specifically, science mapping analysis allow us to uncover

the conceptual, intellectual or social aspect of a research field.
Mainly, science mapping analysis is based on bibliographic networks (Batagelj and Cerinšek 2013) which could be based on co-occurrence, coupling, or direct link. Thus, the main bibliographical networks are co-words, co-citation (at article, author or journal level), coupling (e.g. bibliographical coupling is the most common) and co-authorsips.
Börner et al (2003) purposed a general workflow which was enriched later by Cobo et al. (2011b) to add also quantitative indicators. Therefore, a science mapping analysis based on co-words networks has the following stages: i) detection of research themes, ii) visualize the research themes in a strategic diagram or other kind of visualization, iii) detect the evolution areas, and iv) performance analysis.

4. Software tools to carry out analysis

The bibliometrics analysis could be enhanced with scientific maps representing the relationship among the different actors (authors, institutions, countries, etc.). A map of science is, therefore, a spatial representation of how disciplines, fields, specialties and individual papers or authors are related to one another as shown by their physical proximity and relative locations.
In the last years, a great variety of bibliometric and science mapping software tools have emerged (Moral-Muñoz et al. 2020). In what follows, the most important are listed:

- Bibliometrix / Biblioshiny
- CiteSpace
- SciMAT
- VosViewer

5. Challenges and future directions of science mapping analsysis

Although the above techniques are of great versatility for the analysis of the GRAKO, we must emphasize that it is extremely complex, has multiple heterogeneous and dynamic relationships between its components, and also contains a large volume of data. For this reason, it is necessary to design and develop new models, algorithms, and software tools capable of extracting the hidden knowledge from large volumes of information. In fact, we can determine that the scientific and technological information existing today, and which make up the GRAKO, has the following fundamental characteristics: Volume, Variety and Veracity.

These characteristics are common to the so-called Big Data, which tries to process and analyze large volumes of information from various sources, using techniques based on artificial intelligence. In this sense, the analysis of GRAKO can be supported by artificial intelligence and Big Data techniques, so that the science of science (Fortunato et al., 2018) can process the huge amount of existing information generated in the processes of innovation, development and research, and thus obtain useful knowledge previously unknown in the academic and social context.

Thus, at present, GRAKO knowledge extraction faces the following challenges: i) data source, ii) reproducibility, iii) traceability, representation models, iv) analysis and visualizations tools, v) software tools, vi) knowledge transfer to industry and society.

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273–289

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Re-view of Information Science and Technology*, 37(1), 179–255.

Batagelj, V., & Cerinšek, M. (2013). On bibliographic networks. *Scientometrics*, 96(3), 845–864.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). Science map-ping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). 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.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019.

Hicks, D., Wouters, P., Waltman, L., Rijcke, S. de, & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, 520(Abril).

Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman,

L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, 359(6379), eaao0185.
Lebel, J., & McLean, R. (2018). A better measure of research from the global south. *Nature*, 559(7712), 23–26.
Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de La Información*, 29(1).
Sinatra, R., Wang, D., Deville, P., Song, C., & Barabasi, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312), aaf5239–aaf5239.

Sentiment Analysis, Social Media and Subjective Well-Being

Stefano Maria Iacus

Harvard University – siacus@iq.harvard.edu

Abstract
After reviewing a few basic concepts of sentiment analysis for social media analysis, we present the iSA algorithm which is an unbiased, statistically and computationally efficient method for the estimation of the aggregated distribution of opinions in a set of textual data. We discuss further the problems related to different types of bias that arise in the analysis of social media and an attempt to control for it through statistical methods. Finally, we discuss an application of sentiment analysis that aims at extracting expressions of subjective well-being from Twitter data. In this application we show the impact of the recent COVID-19 pandemic on the derived well-being indicators.

Keywords: sentiment analysis, social media, subjective well-being.

1. Methods in sentiment analysis from social media analysis

Sentiment analysis is, generally speaking, about the task of extracting meaning and/or sentiment or opinions from a corpora of (usually unstructured) text. Although this is not a new topic in textual analysis, it became quite prominent since the advent of social network platforms like Twitter and Facebook, where part of the content is made of text. Nowadays there exist a diversified set of machine learning and statistical techniques that can be useful in the context of sentiment analysis. These include both *unsupervised* methods like: Wordfish (Slapin and Proksch, 2008), topic modelling (Blei et al. 2003), cluster analysis, continuous space word representation, or Word2Vec (Mikolov et al., 2013), and *supervised* methods, such as: Wordscores (Laver et al., 2003), Support Vector Machines (Cortes and Vapnik, 1995), Random Forests (Breiman, 2001) and different families of Artificial (eventually deep) Neural Networks, to mention a few, as well as two recent methods that perform aggregated distribution estimation, i.e., ReadMe (Hopkins and King, 2010) and iSA (Ceron et al., 2016). See also Table 1.

Table 1. Classification of methods with example of candidates.

	Type of Learning Algorithm	
Target of Estimation	**Unsupervised**	**Supervised**
Individual estimation	Corpora approach, NLP, WordFish, topic models, word2vec, clustering methods, etc	SVM, Random Forests, Artificial Neural Network, Deep Learning, WordScores, LLS, etc
Aggregated estimation	aggregation of the above	ReadMe, iSA

At least two communities of scientists work on this topic: the statisticians and the computer scientists, not necessarily with the same goal in mind. Indeed, social media data is big data, and the huge amount of information contained within tweets and Facebook posts cannot be decoded and separated from background noise without the help of appropriate statistical techniques. In our brief review of methods we will try summarize the concepts behind different approaches to sentiment analysis and try to convey the idea that human supervision is an essential step of this process whatever technique is used, especially when dealing with social media data. To rephrase Gary King: nowadays, "social science needs to be computer-assisted, but has to be human-empowered". In this talk we will discuss more in detail why the iSA algorithm is a suitable method of sentiment analysis in the social media context.

2. On capturing expressions of subjective well-being in social media and why

One of the advantages of social networking sites as data sources is that users share their reactions to specific events without the need for the researcher to ask a questions. The researcher need only listen to what users are saying and infer from their texts information about mood and emotions. The expressions of feelings and opinions are spontaneous (to some extent) and not encouraged in any way (although some leveraging has been also observed), hence they are more likely to reflect the actual mindset of their authors, or at least how they want to be perceived. Governments and social scientists have tried to assess the state of wellness of a society using indicators such as gross domestic product (GDP) and employment rates, mostly because they are the easiest quantities that can be measured in an objective way. The basic criticism to this approach is that "*money can't buy happiness*" (although it helps!), i.e., happiness and well-being are more than just economic wealth, is about the quality of relationships, of life, is about feelings, health status, etc. Surveys are the traditional method by which

researchers used to investigate around these aspects of well-being, but unfortunately response rates have declined a lot over time and especially during COVID-19 time. Moreover, as the economists Deaton and Stone (2016) noticed, the order of questions in a survey can influence the answers given. Also the timing and the seasonality may hardly affect the perception of one's well-being and, further, asking about past feelings ("how did you feel last March?") is deeply unconsciously flawed due the way human memory works (Gilbert, 2007): indeed most of us very likely remember how we felt on September 11, 2001 but not the day before.

We compare different methods proposed in the recent literature aiming at constructing subjective well-being indicators based on social media data, namely the "Hedonometer" (Dodds et al., 2011), the "Gross National Happiness Index" (Rossouw and Greyling 2020), the "World Well-Being Project" (Guntuku, 2020) and the "Twitter Subjective Well-Being" (SWB) index (Iacus et al. 2020a, 2020b; Iacus and Porro, 2021, 2022).

All these indicators are clearly affected by different types of bias and we will present a couple of examples in which this bias can be controlled (Iacus et al. 2020b; Spyratos et al, 2020).

2.1. The proposed social media indicator

The proposed Twitter SWB indicator uses sentiment analysis to analyze three different areas of well-being: personal (emotions, life satisfaction, vitality, resilience, functioning), social (trust and belonging, quality of relationships) and *well-being at work* (satisfaction at work). The basic process is as follows. First, the semantic content of Twitter messages is extracted through the manual classification of a sample of texts being classified as either "*negative*", "*neutral*", "*positive*", or "*off-topic*" according to each of the eight dimensions that define the three areas of well-being (emotions, life satisfaction, etc). Then the iSA algorithm estimates the aggregated distribution of positive, negative, neutral and off-topic categories for all tweets, for each of the eight dimensions. Indices are calculated as the proportion of positive sentiment over the total of positive and negative sentiment, disregarding the neutral and off topic categories. The indices are then rescaled to the interval [0-100]% and the final SWB index is the average of these eight indicators.

3. The impact of COVID-19 on well-being

As case study will discuss what has happened to the proposed social media indicators during the Covid-19 pandemic (Carpi et al., 2021). Figure 1 shows the daily oscillations as well as the estimated trends for the two countries Italy and Japan. It turns out that both indexes decreased substantially during the period of observation from 2019 to 2020. In further analysis we will also try to answer

to a few of these questions through a data science approach: *How did people feel about orders to "stay at home"? How did they cope with the deaths of friends and relatives, or the loss of jobs? Were they confident, concerned, or scared about their futures?*

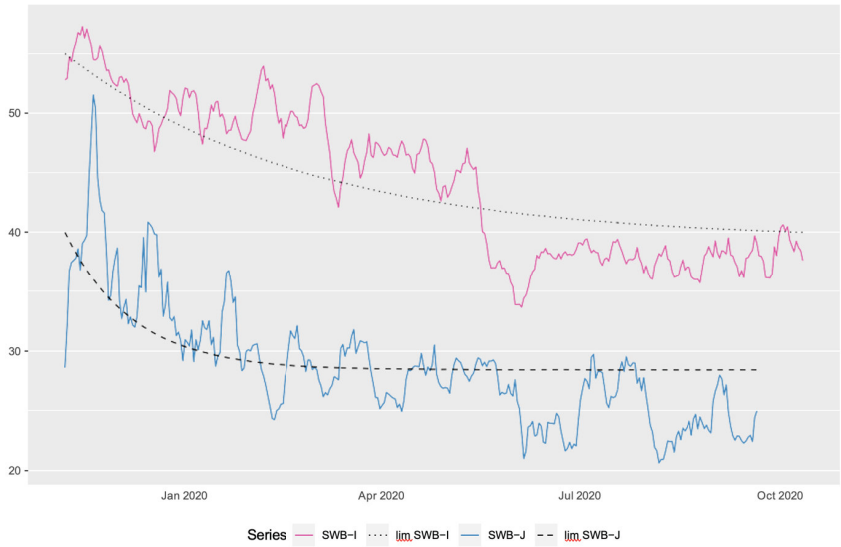


Figure 1. The daily oscillation of SWB-I and SWB-J from November 2019 to 10 October 2020 for Italy and 20 September for Japan, and their trend estimates ("lim SWB-I" and "lim SWB-J"). The vertical axis represents the value of the index on the [0-100]% scale. The total volume of tweets is 13,975,242 for Italy and 12,907,902 for Japan.

References

Blei D.M., Ng A. and Jordan, M. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Breiman L. (2001). Random Forests. *Machine Learning*, 45: 5-32.

Carpi T., Hino A., Iacus S.M. and Porro G. (2021) Twitter Subjective Well-Being Indicator During COVID-19 Pandemic: A Cross-Country Comparative Study. Technical report, https: //arxiv.org/abs/2101.07695.

Ceron A., Curini L. and Iacus, S. M. (2016) iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, 367–368:105–124.

Cortes C. and Vapnik V. (1995), Support-vector networks. *Machine Learning*, 20: 273-297.

Deaton A. and Stone A.A. (2016). *Understanding context effects for a measure of life evaluation: How responses matter.* Oxford Economic Papers, 68(4): 861-870.

Dodds P.S., Harris K.D., Klouman I.M., Bliss C.A. and Danforth C.M. (2011) Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter, *PLoS ONE*, 6(12): e26752.

Gilbert D. (2007). *Stumbling on Happiness*. Vintage.
Guntuku S.C., Sherman G., Stokes D.C., Agarwal A., Seltzer E., Merchant R.M. and Ungar L.H. (2020). Tracking Mental Health and Symptom Mentions on Twitter During COVID-19. *Journal of General Internal Medicine*, 35: 2798-2800.
Hopkins D. and King G. (2010) A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1): 229-247.
Iacus S. M. and Porro G. (2021). *Subjective Well-Being and Social Media*. Boca Raton, FL: CRC Press.
Iacus S.M. and Porro G. (2022). Don't ask, just listen . . . Using social networks to measure subjective well-being. *Significance*, 19: 10-15.
Iacus S.M., Porro G., Salini S. and Siletti E. (2020a) An Italian Composite Subjective Well- Being Index: The Voice of Twitter Users from 2012 to 2017. *Social Indicator Research*, 161: 471- 489.
Iacus S.M., Porro G., Salini, S. and Siletti E. (2020b) Subjective Well-being, Social Media Data and Sampling Bias: a Tentative Solution. *Journal of Official Statistics*, 36(2): 315-338.
Laver M., Benoit K. and Garry J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *The American Political Science Review*, (97)2: 311-331.
Mikolov T., Sutskever I., Chen, K., Corrado G. and Dean J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 3111-3119.
Rossouw S. and Greyling T. (2020). Big data and happiness. In Zimmermann, K.F.. editor, *Handbook of Labor, Human Resources and Population Economics*. Springer, 1-35.
Slapin J.B. and Proksch S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3): 705-772.
Spyratos S., Vespe M., Natale F., Iacus S. M. and Santamaria C. (2020). Explaining the travelling behaviour of migrants using Facebook audience estimates. *PloS one*, 15(9): e0238947.

Linguistic Resources for Corpus Processing: the ATISHS project[1]

Max Silberztein

Université de Franche-Comté – max.silberztein@univ-fcomte.fr

Abstract
Nowadays, most corpus processor software applications used in the Digital Humanities analyze texts as if they were constituted by sequences of graphical wordforms. However, users of these software who work in the social sciences and the humanities are looking for information expressed by units of meaning. It is therefore necessary to establish a link between the graphical wordforms that constitute the text files and the units of meaning that are of interest to the users of these software. We show that creating this link requires precise linguistic methods and resources. We present the linguistic resources developed with the NooJ linguistic platform and demonstrate how they are used by the new ATISHS corpus processor to provide more reliable statistical analyses for the digital humanities.

Keywords: Corpus Processing, Digital Humanities, Linguistic resources.

1. Introduction

Nowadays, most researchers in the humanities and the social sciences explore and study their corpora with corpus processor software applications such as Alceste[2], Hyperbase[3], IRaMuTeQ[4], Lexico[5], Sketch Engine[6], TXM[7], Word2Vec[8]. All these software applications process text files as sequences of graphical wordforms, on the assumption that these graphical wordforms represent the pieces of information useful to their users:[9] their concordances display

[1] This research was supported by the project "Analyseur de Textes Innovant pour les Sciences Humaines et Sociales" [A new Text Analyzer for the Social Sciences and the Humanities] from the Region of Franche-Comté, see: http://www.nooj-association.org/atishs.html.
[2] Reinert, 1999.
[3] Brunet, 2011.
[4] Loubère et al., 2014.
[5] Lamalle et al., 2002.
[6] Kilgarriff et al., 2014.
[7] Heiden, 2010.
[8] Church, 2017.
[9] By (graphical) wordforms, we mean contiguous sequences of letter characters delimited by non-letters (a.k.a. delimiters). For instance, "tomorrow" and "stepfather" are two wordforms, whereas the similar sequences "next day" and "father-in-law" contain respectively two and three wordforms.

wordforms in context; their search engines compile indices of wordforms in order to answer their users' queries (themselves segmented as sequences of wordforms); the word clouds they provide display wordforms; their statistical analyses detect interesting frequencies of wordforms (or collocations of wordforms), etc.

However, wordforms almost never correspond to the units of meaning — concepts, entities, predicates and relations — that are sought by the researchers in social sciences and the humanities who study their corpus.

In the NooJ framework,[10] units of meanings are elements of a language's vocabulary and are named "Atomic Linguistic Units" (ALUs).[11] By definition, ALUs are *atomic*, *i.e.* not to be analyzed: locutors have learned them by heart, learners of second-language cannot invent them, and linguistic computer software can only find their properties by consulting dictionaries.

Although linguists have not always agreed on how to characterize ALUs, applications such as Machine Translation or Second-Language teaching make it easier to understand what vocabulary elements are. For example, no English speaker would try to guess how to say "tank top" in French by translating "tank" [réservoir] and "top" [dessus]: *tank top* must be translated as one single French noun (*débardeur*): tank top is an ALU. Reciprocally, no French person should try to translate "pomme de terre" [potato] in English by translating "pomme" [apple], "de" [of] and "terre" [earth]: *pomme de terre* is an ALU, and MT software should directly link the sequence "pomme de terre" to "potato", by just looking up a bilingual dictionary.

Conversely, people who learn French do not need to learn the meaning of the verbs "redormir", "remanger", "recréer", etc. They just need to learn the base verbs *dormir* [to sleep], *manger* [to eat] and *créer* [to create], and know that the prefix "re-" can be added in front of any French verb to express repetition: "re-" is a French ALU.

[10] In the following, we will use the NooJ notation to display dictionary, grammars and grammar rules, see (Silberztein 2016). NooJ is a free and open-source linguistic development environment used to formalize linguistic phenomena (from orthography to semantics) for over 30 languages, as a corpus processor used in the Digital Humanities, as well as as a linguistic engine to develop NLP software applications. To download NooJ, its manual, its linguistic resources as well as access tutorials and references, see: http://www.nooj-association.org.

[11] In NooJ's linguistic engine, ALUs are represented as annotations such as: <eaten, eat, VERB +Transitive +N0Anim +N1Conc +Pastparticiple>, where graphical wordforms (e.g. "eaten") are associated with a lexical entry (e.g. "eat"), a category (e.g. "VERB") and a list of properties (e.g. +Transitive, +N0Anim, +N1Conc and +Pastparticiple).

Therefore, there has to be an intermediary step during which corpus processors link the graphical wordforms that occur in the text files to the ALUs that are actually of interest to their users. In the following, we show why linguistic methods and resources are needed, and how they are used in the ATISHS software.

2. Spelling

Obviously, graphical wordforms are defined by their spelling. However, spelling is not always fixed. For example, in the Meditext XIVth century corpus of French texts, the term *seigneur* [lord] has over fifty spelling variants. In order to offer historians and medievalists the means to explore and analyze their corpus, a corpus processor must therefore identify the term *seigneur* with all its spelling variants. In the PALM exploration software,[12] this mapping is performed by using the following NooJ grammar.

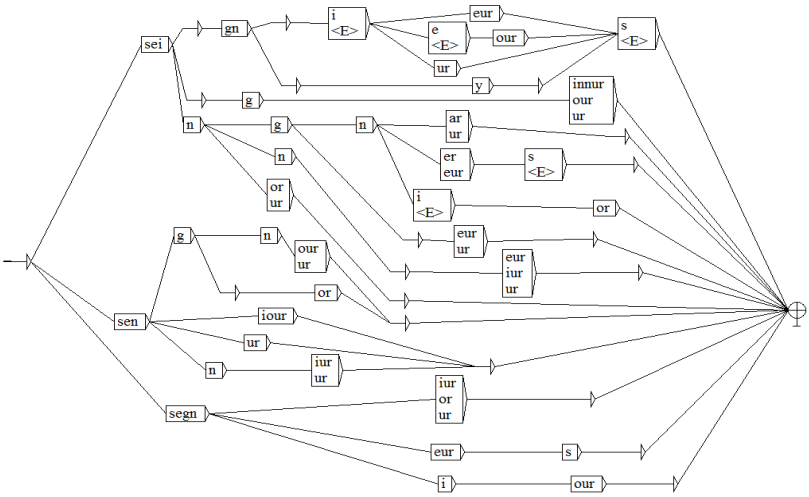


Figure 1. Grammar that represents all spelling variants of the term seigneur [lord]

The grammar in figure 1 recognizes any wordform that can be spelled by following a path that starts at the initial node (the horizontal arrow on the left of the graph), follows any connection, and ultimately reaches the terminal node (the

[12] Aouini 2018.

target on the right of the graph). Using this grammar, the PALM corpus processor finds all the occurrences of the term *seigneur* [lord] in the corpus — whatever their spelling — and can therefore produce reliable statistical analyses of this term: characterize its frequency in sub-corpora, find interesting collocations with other terms such as *vassal*, *liege*, etc., study the evolution of its use in different regions, etc.

Today, spelling is not much more reliable in the texts typically used by researchers in the social sciences: tweets, blogs, text messages, emails etc. But even in better spelled texts such as newspaper articles or works of literature,[13] spelling variants are very frequent. There is systematic hesitation in the spelling of words of foreign origin (e.g. *yoghurt* vs. *yogurt*; *oenolo*gy vs. *enology*), in the use of space and hyphenation (e.g. *audiovisual*, *audio-visual*, *audio visual*), capitalization (e.g., Middle Ages vs. middle ages),[14] mute letters (e.g., *judgement* vs. *judgment*), double letters (e.g. *traveled* vs. *travelled*), word endings (e.g. *program* vs *programme*)[15], regional spelling variants (e.g., *harbour* vs. *harbor*), etc. Acronyms, synonyms and abbreviations add further complexity to the mere recognition of terms and proper names (e.g. *US Republican Party* vs. *Grand Old Party* vs. *GOP*, the Conservative Party), etc.

By ignoring variants of a term they process (find, count, analyze, compare etc.), corpus processors systematically underestimate its frequency and thus its relevance in texts. For example, a Google search only finds 70 million occurrences for the query "audio visual," whereas it finds over 700 million occurrences for the query "audiovisual." With only 10% of the results found, we are far from the 95% recall rate typically claimed by most Natural Language Processing applications used in the Digital Humanities.

[13] (Mathieu-Colas 1990) demonstrated that over 5% of the spelling of the entries of the three French reference dictionaries (Grand Dictionnaire Encyclopédique Larousse, Grand Robert and Trésor de la Langue Française) were inconsistent.

[14] In order to fix the problem of massive case variation, most computer software are case-insensitive. However, ignoring case makes these software produce a number of false positives, such as finding March (the month) when looking for the verb to march, Polish (from Poland) for the verb to polish, etc. More systematically, corpus processors that use case-insensitive tokenizers cannot distinguish acronyms from proper names and common nouns, e.g. RAM (Random Access Memory) vs. Ram (variant of Rāma) vs. a ram (the sheep).

[15] In French, plural endings (-s) are mute. As a consequence, there is a systematic variation between singular and plural components for tens of thousands of terms, such as a flight simulator (un simulateur de vol or un simulateur de vols?), a citrus juicer (un presse-agrume or un presse-agrumes?), etc.

To offer users of corpora truly reliable analyses, a corpus processor must therefore take terms' variations into account. Variations can be described either by grammars such as the one in figure 1 or by dictionaries. For example, figure 2 shows a dictionary sample that describes the term tsar and its spelling variants:

```
tsar,N
csar,tsar,N
czar,tsar,N
tzar,tsar,N
```

Figure 2. The lexical entry tsar and its three variants

In this dictionary, the wordforms "csar", "czar" and "tzar" are linked to the lexical entry "tsar." By accessing this dictionary, a corpus processor can collect all variants of the term tsar before counting its occurrences: it will thus get a better recall rate.

3. Inflectional Morphology

Some tools used in the Digital Humanities do not perform any morphological analysis of wordforms before performing sophisticated statistical computations on their frequencies: they treat wordforms such as the following ones as independent units: *march*, *marches*, *marching*, *marched*. The authors of these software justify this policy. For example, for Lexico, (Lamalle et al., 2002) note that in left-wing texts of the 1970s, one uses the plural term *libertés* in the context of defending social rights to housing, education, etc., whereas the singular term *liberté* occurs in right-wing texts to designate the concept of entrepreneurial freedom. These are two different concepts, and their occurrences should not be unified nor their frequencies aggregated.

This argument is relevant: there are indeed abstract nouns that do not take a plural form in a semantically transparent way. However, over 95% of abstract, concrete, animated and human nouns can be put in the plural with no difference in meaning other than their number:

(Abstract noun): *One demonstration* ⟶ *two demonstrations*
(Concrete noun): *One table* ⟶ *two tables*
(Animated noun): *One giraffe* ⟶ *two giraffes*
(Human noun): *One baker* ⟶ *two bakers*

Treating occurrences of the two wordforms "demonstration" and "demonstrations" as unrelated units, for instance, would be problematic for a historian or a sociologist who seeks to evaluate the number of conflicts in a

country by exploring a newspaper archive. By not adding the frequencies of these two wordforms, a statistical analyzer would underestimate the overall frequency of this term, hence it might miss its importance and relevance.
A better solution is to lemmatize wordforms that can be put in the plural in a transparent way, and still treat as independent entries those whose meaning is different in the singular and in the plural form. That can be implemented by a dictionary such as the following:

```
liberty,N+Singular
liberties,N+Plural
demonstration,N+FLX=TABLE


TABLE = <E>/Singular | s/Plural;
```

Figure 3. Three lexical entries and one inflectional paradigm

The dictionary in figure 3 formalizes the fact that *liberty* (= freedom) and *liberties* (= breach of social convention) are two independent terms, whereas the term *demonstration* is associated with two wordforms, generated by the inflectional paradigm rule TABLE (i.e. singular, takes an "s" in the plural).
A more general argument used to not lemmatize wordforms is that the process of lemmatization intrinsically loses information that might be relevant. For example, for a literary study, a psychological study or a semantic analysis, the difference in meaning carried by the wordforms *ate* and *eaten* could be significant, as these two wordforms have different narrative and logical values: *Joe has eaten already* implies that he has finished eating (he is probably not hungry anymore), which cannot be inferred by *Joe ate an apple yesterday*.
However, there are also crucial narrative and logical differences in the three following sentences:
Eat your soup! I only eat in kosher restaurants. Next week, we'll eat in an Italian restaurant.
In the first sentence, the wordform *e*at represents an order, to be followed in the near future; in the second, it represents a past and present habit of the narrator; in the third, it represents an activity that will happen in the future. In these three sentences, the wordform *eat* has three different narrative values: it does not make more sense to aggregate them together while separating the wordforms *ate* and *eaten*.[16] Distinguishing the wordforms *eaten* and *ate* while aggregating the different values of the wordform *eat* is therefore not consistent.
A more consistent approach would be to process units that are not the mere

[16] Reciprocally, note that the two following sentences are synonymous: "Have you eaten yet?" = "Did you eat yet?".

graphical wordforms, but the ALUs themselves: <ate=eat,VERB+Preterit>, <eaten=eat,VERB+Pastparticiple>, <eat=eat,VERB+Infinitive>, <eat=eat,VERB+Imperative> and <eat=eat,VERB+Present>. That way, the user can choose to perform statistical analyses on the wordforms (before the equal sign) or on the lemmas (after the equal sign), according to their need.

4. Stochastic vs. Linguistic Lemmatization

Most software tools used in the digital humanities do offer some lemmatization functionalities, provided by stochastic lemmatizers such as *Gate*,[17] *NLTK*[18] or *TreeTagger*.[19] Stochastic lemmatizers tag the user's text by comparing the contexts of its wordforms with the ones found in a reference training corpus.[20] Unfortunately, these "reference" corpora contain a large number of mistakes. For example, (Taylor, 2003) shows that the Penn Treebank contains many mistakes such as "Battle-tested" and "Japanese" tagged as proper names; (Silberztein, 2018) shows that the Open American National Corpus (OANC), tagged with *Annie*,[21] contains over 11,019 impossible tags such as *abbreviate* tagged as a noun, *about* as an adjective, *anomaly* as an adverb, etc. In the COCA corpus, many typos such as "accrossthe" are tagged as nouns, the wordform "that" is incorrectly tagged as a conjunction in "The one that is most concerned," as a determiner in "My contention, broadly stated, is that we ought to concern ourselves...", as an adverb in "Ellen Ashdown pointed out that many humanities...", etc.[22]

~~Battle-tested/SingularProperName~~ ~~Japanese/SingularProperName~~ industrial/Adjective managers/PluralNoun here/Adverb always/Adverb buck/Verb up/Preposition nervous/Adjective newcomers/PluralNoun with/Preposition the/Determiner tale/SingularNoun of/Preposition the/Determiner first/Adjective of/Preposition their/PossessivePronoun countrymen/PluralNoun to/to visit/Verb Mexico/SingularProperName,/, ~~a/Determiner~~ ~~boatload/SingularNoun~~ ~~of/Preposition~~ ~~samurai/PluralNoun~~ ~~warriors/PluralNoun~~ ~~blown/PastParticiple~~ ~~ashore/Adverb~~ ~~375/Number~~ ~~years/PluralNoun~~ ~~ago/Adverb~~ ./. ~~From/Preposition~~ ~~the/Determiner~~ ~~beginning/SingularNoun~~ ,/, ~~it/PersonalPronoun~~ ~~took/PastParticiple~~ a/Determiner man/SingularNoun with/Preposition extraordinary/Adjective qualities/PluralNoun ~~to/TO~~ succeed/Verb in/Preposition Mexico/SingularProperName ,/, "/" says/VerbPresent3rdSingular Kimihide/SingularProperName Takimura/ SingularProperName ,/, president/SingularNoun of/Preposition ~~Mitsui/PluralNoun~~ ~~group/SingularNoun~~ 's/Possessive ~~Kensetsu/SingularProperName~~ ~~Engineering/SingularProperName~~ ~~Inc./SingularProperName~~ unit/SingularNoun ./.

Figure 4. Mistakes in the Penn Treebank are crossed out; unrecognized multiword units are underlined.

[17] Cunningham, 2002.
[18] Loper et alii, 2002.
[19] Schmid 1994.
[20] Such as, for English: the Penn Treebank, the Corpus of Contemporary American English (COCA), the Open American National Corpus (OANC), etc.
[21] Annie is a Gate plugin.
[22] In figure 4, I have replaced Penn Treebank's tag codes (e.g. "NNP") with a more readable form (e.g. "SingularProperName"). The fact that these training corpora contain large number of mistakes is well known and has been the subject of many publications, e.g. (Dickinson 2012), (Green & Manning 2010), (Volokh & Neumann 2011), etc.

We should not expect Natural Language Processing software that use unreliable references to produce reliable results.

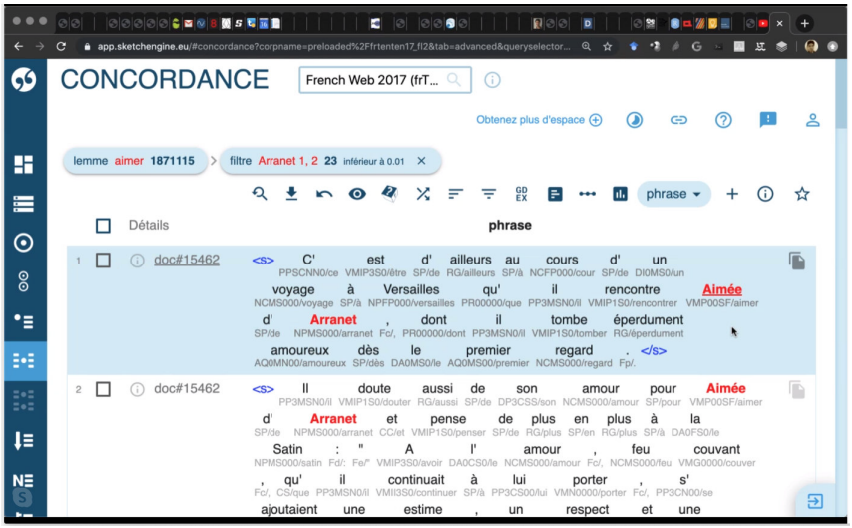


Figure 5. Tagging mistakes in Sketch Engine

For instance, here is a French sentence[23] tagged by Sketch Engine:
C'est d'ailleurs au cours d'un voyage à Versailles qu'il rencontre Aimée d'Arranet, dont il tombe éperdument amoureux dès le premier regard.
[By the way, it is during a trip to Versailles that he meets Aimée d'Arranet whom he instantly falls madly in love with].
In this single sentence, *cours* has been incorrectly tagged as "NCFP000/cour", i.e. a plural form of the feminine noun *cour* [courtyard], *Aimée* as "VMP00SF/aimer", i.e. a past participle form of the verb *aimer* [loved], *rencontre* as "VMIP1S0/rencontrer", i.e. a first person singular form of the verb *rencontrer* and *tombe* as "VMIP1S0/tomber", i.e. a first person singular form of the verb *tomber*. Moreover, the tagged text does not represent the fact that "d'ailleurs" [by the way] is an adverb rather than a sequence of a preposition and a locative adverb, "au cours de" [during] a preposition (no relation whatsoever with *courtyards*), "tomber amoureux" [fall in love] a frozen expression and "dès le premier regard" [instantly] an adverb. Producing 6

[23] From the fictional character Nicolas Le Floch's biography, a series of detective novels that take place in the XVIIIth century, by Jean-François Parot (1946-2018).

incorrect tags and ignoring 4 multiword units in a single 26-wordforms-long sentence cannot be good enough for any application in the digital humanities. For instance, if a linguist wants to use a corpus to study locative nouns, they will get many false positives such as *cour* [courtyard], that will produce higher frequencies and incorrect collocations. If a researcher in literature performs a stylistic analysis of this text, they will grossly underestimate the number of an author' use of concessive adverbs such as *d'ailleurs* [by the way]. If they perform a narrative analysis of the text, they will greatly overestimate the author's use of verbs conjugated in the first person, etc.

The principle of using reference training corpora to analyze texts, while standard in today's NLP, has several shortcomings:[24]

— Natural languages contain an infinite set of potential sentences. Therefore, however large a training corpus is,[25] it will never be big enough to cover all the linguistic contexts one might encounter when analyzing texts as varied as the ones used by historians (e.g. XIVth century corpus), psychologists (e.g. children interviews), researchers in literature studies (e.g. theater plays, poems), sociologists (e.g. polls), etc.;

— Analyzers based on training corpora compare the context of the wordforms in the text they are analizing with the contexts of the same wordforms in the training corpus. These contexts are typically defined as sequences of a few wordforms. However, texts written in natural languages do not have a linear structure: most ALUs may occur before of after any sequence of ALU categories. For instance, the wordform "sees" might be followed by an adjective (e.g. *Joe sees pretty flowers*), a determiner (e.g. *Jane sees some flowers*), a noun (*He sees flowers*), a conjunction (e.g. *Nobody sees but you*), a preposition (e.g. *She sees behind herself*), a pronoun (*Everyone sees it*), an adverb (e.g. *She sees very well*). In these conditions, it is not surprising that analyzers that compute results based on these limited contexts produce mistakes.

In fact, the excellent metrics obtained by stochastic analyzers (typically 95% recall and precision) can easily be explained by the fact that their tagset is limited (e.g. no difference between abstract nouns and concrete nouns; no distinction between the various meanings of a given verb; not even any distinction between intransitive and transitive verbs; etc.).

The linguistic approach constitutes a more reliable approach: carefully construct handcrafted dictionaries so that they do not contain typos such as "accrossthe =

[24] Silberztein 2016 lists ten shortcomings of stochastic approaches to Natural Language Processing.

[25] Several projects aim at constructing larger and larger training corpora, such as LDC's English gigaword dataset, cf. https://catalog.ldc.upenn.edu/LDC2003T05 and https://deepai.org/dataset/gigaword (4 million articles).

Noun", label each lexical entry to avoid incorrect analyses such as "anomaly = Adverb," do not ignore multiword units such as "a boatload of = Determiner," and describe the inflectional paradigm of each lexical entry properly:

```
abbreviate,VERB+FLX=LIVE+Transitive
about,PREPOSITION
anomaly,NOUN+FLX=ABILITY+Abstract
a boatload of,DETERMINER
eat,VERB+FLX=EAT+Transitive
many,DETERMINER+Plural

LIVE = <E>/INF | s/3+s | <B>ing/Gerund | d/Preterit | d/PastParticiple
ABILITY = <E>/Singular | <B>ies/Plural
EAT = <E>/INF | s/3+s | ing/Gerund | <B3>ate/Preterit | en/PastParticiple
```

Figure 6. Lexical entries and inflectional paradigms in an error-free dictionary

By looking up a dictionary that is complete and does not contain any mistake, a corpus processor can lemmatize ALUs with a perfect recall rate and a perfect precision:
— the recall rate would be perfect, as if an ALU were to be missing its proper analysis, a lexicographer could just add it to the dictionary instantly;
— the precision would be perfect, as if an incorrect analysis were to be found, a lexicographer could correct the corresponding dictionary instantly.[26]
Such a dictionary should contain all the elements of the standard vocabulary, which for English amounts to less than 100,000 simple words and 250,000 multiword units. Proper names could be recognized by accessing specialized dictionaries such as JRC-Names.[27] Technical and specialized terms could be recognized by accessing dictionaries adapted to the corpus domain such as Medline for medical texts.[28]

5. Derivation

As in the aforementioned case study of a sociologist who seeks to evaluate the state of social peace by analyzing archives of a daily newspaper, simply lemmatizing the wordform *demonstration* would not be enough: the corpus processor should take into account all its derived forms to find mentions of this

[26] One must not confuse tagging mistakes (such as tagging "anomaly" as an adverb) and ambiguities (involved when a wordform corresponds to multiple potential tags). We discuss how to represent homography and solve wordform ambiguities in section 7.
[27] See https://ec.europa.eu/jrc/en/language-technologies/jrc-names.
[28] See https://www.nibib.nih.gov/content/medical-dictionary-medlineplus.

term in the following sentences:
*... **Demonstrators** shouted against police brutality and racism and some held signs against discrimination ...*
https://www.aa.com.tr/en/europe/demonstrators-in-france-protest-george-floyd-s-death/1867169
*... Wearing face masks, waving black flags and keeping two yards apart, thousands of Israelis **demonstrated** against prime minister Benjamin Netanyahu ...*
https://www.reuters.com/article/us-health-coronavirus-israel-protest/anti-netanyahu-rally-draws-thousands-under-coronavirus-curbs-idUSKBN2210S4
*... Parents to **redemonstrate** against air pollution, demanding officials to take immediate and efficient actions and for the well-being of children...*
https://mongolia.gogo.mn/r/156657
Morphological families can be described either using dictionaries associated with morphological rules such as the ones in figures 3 and 6, or by grammars such as the following:

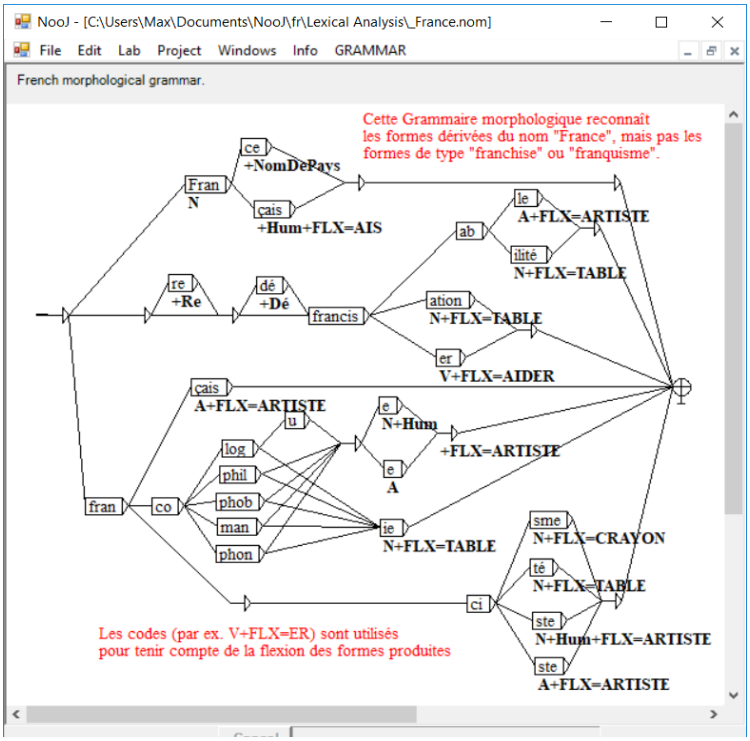


Figure 7. Family of wordforms derived from the ALU France

The grammar in figure 7 recognizes the wordforms *Français* [Frenchman] *franciser* [to frenchify], *francophone*, *francophobe* [Francophobic], etc., as well

as all their inflected forms (e.g. *Françaises*, *franciserons*), which amount to more than three hundred wordforms.

Users of corpus processors that have access to this grammar will be able to analyze the importance of France's cultural or political impact in newspapers such as *Jeune Afrique* or *Le matin d'Algérie* and analyze its use in relation to other concepts, detect whether it occurs in positive or negative sentences, find collocations with interesting terms, find out if it is more present in left-wing or right-wing political articles, characterize the speeches of certain political figures, etc.

Corpus processors that do not have access to this description will underestimate the frequency of France-related topics in a corpus. For example, in the archives of the newspaper *Le Monde*, 2006, the wordform *France* occurs 38,000 times, whereas its derived forms occur more than 60,000 times.

Some corpus processors use root-based morphological parsers to overcome the fact that they do not access any derivational description. For example, Alceste[29] and Hyperbase[30] use orthographic rules such as "am.*" to compute derived wordforms of *amour* [a love]. Unfortunately, a consequence of this approach is that *amour* gets linked to *amphithéâtre* [amphitheater], *ample*, *amplifier* [to amplify], *amuser* [to amuse], etc. and not to any of the conjugated forms of *aimer* [to love].

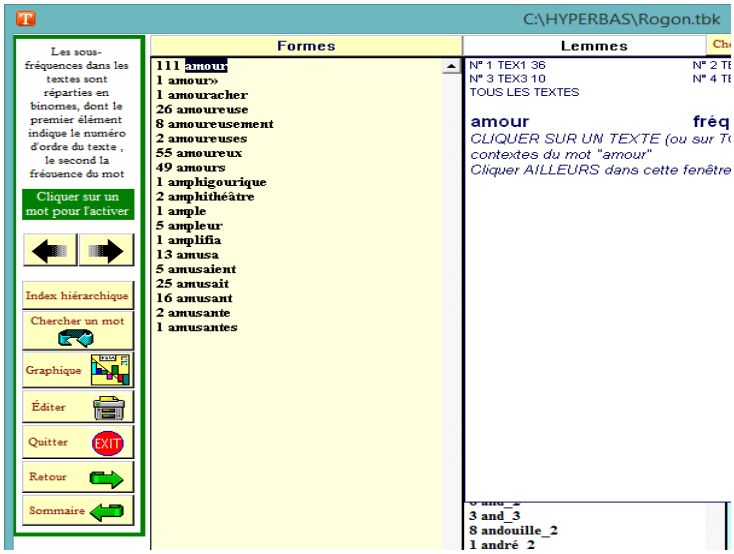


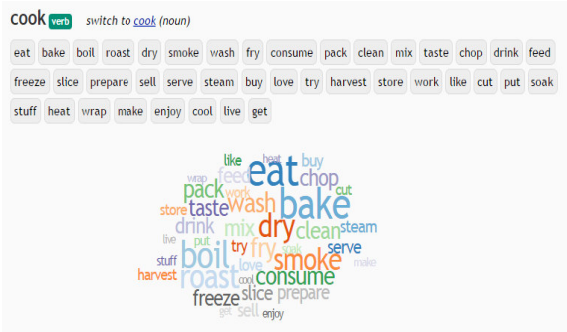
Figure 8. List of wordforms proposed by Hyperbase for the requested term "amour" [love]

[29] Reinert 1999.
[30] Brunet, 2011.

Generally, orthographical parsers are not reliable. For instance, if users are looking for all France-related sentences in newspaper archives, selecting all the wordforms that match the regular expression "fran.*" would produce a large number of false positives, such as *franchir* [to cross (a river)], *frangipane* [almond cake], *franchise* [Honesty], while missing terms such as *refranciser* [re-frenchify].

6. Word clouds and semantic networks

Figure 9. A word cloud for the wordform cook provided by Sketch Engine.



To help users study a concept in their text corpora, some corpus processors display a cloud of related words such as the one in figure 9. However, most of the words shown in these clouds are not particularly interesting; for example, the words *dry*, *pack*, *clean*, *consume*, *get*, *try*, *pack*, *work*, etc. do not add any specific context to the concept of cooking, and there is no benefit in presenting grammatical words such as *get*, *like* or *make* to the users. A better approach would be to let users control what terms should be looked for, using precisely defined queries such as the following:[31]

```
COOK = <bake> | <boil> | <brew> | <broil> | <cook> | <cuisine> |
<fry> | <grill> | <roast> | <simmer> | <steam> | <stew> | <toast>
```

Each user should be able to define lexical fields according to their specific needs. For instance, a psychologist might want to aggregate occurrences of the verbs to love and to hate when analyzing the emotivity of a patient, whereas a political scientist who analyzes speeches of various political figures might want to oppose

[31] In NooJ, symbols in angle brackets represent sets of morphological-related wordforms. For instance, <bake> = bake | bakes | baking | baked | baker | bakers | bakery | bakeries. Users can add constraints to symbols; for instance, <bake,V> represents only the verbal forms bake, bakes, baking and baked.

these verbs. For an analysis of tourism-related brochures, the COOK lexical field might include terms such as *coffee shop, cooking class, diner, fast-food, gastronomy, gourmet store, restaurant*; for an analysis of dietary-related medical articles, it could include the terms *anorexia, appetite, bulimia, obesity, snacking*; for yet other studies, *breakfast, dinner, kitchen, lunch, oven, pot*, etc.

7. Homography and lexical ambiguity

In the previous sections, we have seen that recognizing a certain ALU often requires the corpus processor to take into account multiple wordforms. The reverse is also true: most wordforms potentially correspond to multiple ALUs. For example, the wordform *card* has a dozen unrelated meanings (*business card, credit card, green card*, etc.); the wordform *plant* too (*anchor plant, hay plant, nuclear plant, packing plant, pot plant*, etc.), the wordform *house* too (*acid house, auction house, country house, shotgun house, white house*, etc.).

Homography is an important characteristic of natural languages: we estimate that each simple French simple noun has over five meanings on average.[32] Moreover, the more frequent a wordform is, the more meanings it corresponds to. For instance, the French noun "carte" [board, chip, card, pass, map, etc.] correspond to over 30 ALUs (i.e. different meanings).

If a wordform corresponds to multiple unrelated meanings, processing it as a single unit is non sensical: what can a political scientist deduce from a statistical analysis that reveals that the wordform "plant" had a peak of frequency in January 2021? Was this term related to an FBI agent in a covert operation (a "plant")? to the legalization of cannabis, or to a power outage due to an electrical plant failure?

A corpus processor that processes texts at the wordform level might detect that the wordform *matter* has an abnormally high frequency in a certain article of a newspaper and classify it in the scientific section (*gray matter, dark matter*, etc.), whereas this wordform would just appear in adverbs such as "as a matter of fact," "for that matter," "in these matters," etc.

In reality, when they analyze their corpus, most researchers in the humanities and the social sciences are looking for (specific) units of meanings rather than (highly ambiguous) graphical wordforms. Corpus processors must therefore be able to infer the units of meanings from the wordforms that occur in texts.

Lexical disambiguation is a complex problem that we cannot develop here;

[32] Based on the size of the relative size of the French DELAS (35,000 simple nouns) and the DELAC (200,000 compound nouns) dictionaries, cf. (Courtois, Silberztein 1990). Some French nouns have over thirty meanings, for instance *carte*: *carte de crédit* [credit card], *carte routière* [road map], *carte postale* [postcard], *carte-mère* [mother-board], etc. The DELA systems of dictionaries describe natural languages' standard vocabularies: if one adds to them specialized and technical terms, this ratio will grow significantly.

(Silberztein 2010) showed how verbs meaning disambiguation can be achieved using detailed dictionaries associated with refined syntactico-semantic grammars.[33] However, to retrieve terms — that constitute most of the needs of corpus users in the social sciences — a much simpler algorithm can be implemented. Its principle is that, instead of processing *card* as an highly ambiguous wordform, it is better to process its multiword units independently: *business card, credit card, green card* etc. These explicit forms are listed in dictionaries such as DELACs.[34]

Most multiword units can be abbreviated; for instance, a *charitable organization* is an organization; the *ministry of education* is a ministry, the *COVID pandemic* is a pandemic, etc. In practice, multiword units are often abbreviated, as can be seen in:

> *... A third of doctors have asked for mental health support during the Covid pandemic, according to one of the UK's most senior medical organisations. The Royal College of Physicians (RCP) in Wales said Covid had thrown the NHS workforce into "sharp relief". One junior doctor Dr Emma Rengasamy said working through the pandemic was a "baptism of fire" (BBC News, 16 March 2021).*

To recognize ALUs even when abbreviated, a corpus processor could access grammar rules such as the following:

```
COVID = (COVID | COVID-19 | coronavirus) (crisis | epidemic |
pandemic | surge)
```

When this grammar trigger a match in a text, the corpus processor would "activate" this ALU. After being activated, all its subsequent abbreviated occurrences (e.g. *crisis*, or *pandemic*) would then be aggregated and linked to the "COVID" ALU, until of course another ALU (e.g. *migrant crisis*) gets activated.

8. Conclusion and perspectives

Today, more and more researchers in the humanities and the social sciences use corpus processors to explore and analyze their corpora. They are looking for units of meanings rather than graphical wordforms. We have shown that

[33] Dubois & Dubois-Charlier's dictionary *Les Verbes Français* (LVF) contains over 20,000 lexical entries, cf. (François et alii 2007). Each entry corresponds to a specific meaning of a verb, associated with a syntactico-semantic characterization of its context. (Silberztein 2010) describes how this dictionary was used to automatically characterize and retrieve each of the five specific meanings of the verb *abriter* [to house, to shield, to host, to shelter, to hide oneself] in *Le Monde diplomatique* (2002).

[34] The French DELAC dictionary contains multiword units and was developed by (Silberztein, 1990).

exhaustive precise linguistic resources and methods can be used to link the wordforms that occur in the text files to the units of meaning that are of interest to the users.
— taking into account a large number of spelling, inflectional, derivational and synonymous variants of terms will lead to a much better recall rate;
— distinguishing between a wordform's different meanings will lead to a better level of precision.

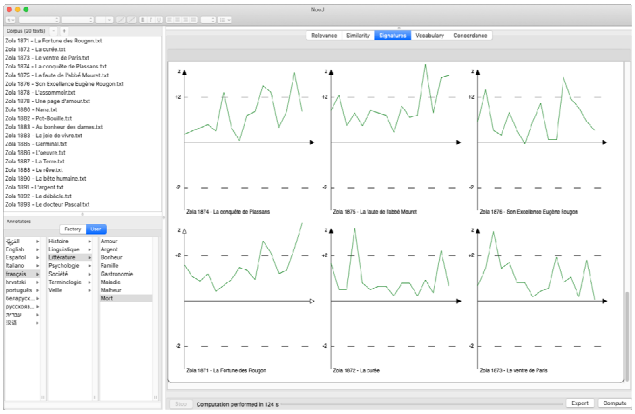


Figure 10. ATISHS' Standard-Score Analysis of the Lexical Field "mort" [death] in Zola's novels.

The ATISHS corpus processor[35] is being developed to provide users with all the statistical tools usually provided by existing corpus processors, applying these tools to units of meaning (ALUs) rather than graphical wordforms. For instance, figure 10 displays a standard score analysis for the occurrences of the lexical field *mort* [death] in the novels of Zola's Rougon-Macquart series. Here is the definition of the lexical field:

```
mort = (<déceder> | <mourir> | <périr> | <suicider>) |
      <cadavre> | <cimetière> | <crémation> | <crématorium>
| <deuil> | <enterrement> | funérailles | <DET> (<A> | <E>)
<tombe> | <veuf> |
      <fatal> | <mortel> ;
```

ATISHS processes lexical fields defined by linguistic resources developed with

[35] cf. http://www.nooj-association.org/atishs.html. All linguistic resources used by ATISHS have been developed using the NooJ linguistic development environmnent, cf. http://www.nooj-association.org/.

the NooJ platform, that include lexicons of terms associated with their morphological and spelling variants, abbreviations and acronyms and synonyms, morphological grammars that represent their inflected and derived forms, as well as syntactic grammars that describe syntactico-semantic contexts used to eliminate false positives.

Although the linguistic approach has an initial high cost since it requires the development and maintenance of linguistic resources, we believe that it is the only valid approach if one wants to offer researchers in the humanities and the social sciences reliable tools to analyze their corpus.

References

Aouini M. (2018). *Approche multi-niveaux pour l'analyse des données textuelles non-standardisées: corpus de textes en moyen français* (Doctoral dissertation, Bourgogne Franche-Comté University).

Brunet E. (2011). Hyperbase : Logiciel hypertexte pour le traitement documentaire et statistiques des corpus textuels : http://hyperbase.unice.fr/hyperbase/doc/manuel.pdf

Church K. W. (2017). Word2Vec. *Natural Language Engineering*, 23(1), 155-162.

Courtois B. and Silberztein M. (1990). Dictionnaires électroniques du français. *Langue française*, 87(1), 3-4.

Cunningham H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2), 223-254.

Dickinson M. and Ledbetter S. (2012, May). Annotating Errors in a Hungarian Learner Corpus. In *LREC* (pp. 1659-1664).

François J., Le Pesant D. and Leeman, D. (2007). Présentation de la classification des Verbes français de Jean Dubois et Françoise Dubois-Charlier. *Langue française*, (1), 3-19.

Green S. and Manning C. D. (2010, August). Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010) (pp. 394-402).

Heiden S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *24th Pacific Asia conference on language, information and computation* (Vol. 2, No. 3, pp. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., ... and Suchomel V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7-36.

Lamalle C. and Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, 403-412.

Loper E. and Bird S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loubère L. and Ratinaud P. (2014). Documentation IRaMuTeQ. Download from: *http://www. iramuteq. org/documentation/fichiers/documentation_19_02_2014. pdf*.

Mathieu-Colas M. (1990). Orthographe et Informatique: établissement d'un dictionnaire électronique des variantes graphiques. *Langue française*, (87), 104-111.

Reinert M. (1999). Quelques interrogations à propos de l'" objet" d'une analyse de discours de type statistique et de la réponse" Alceste". *Langage & société*, 90(1), 57-70.
Schmid H. (1994). TreeTagger-a language independent part-of-speech tagger. Cf. : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
Silberztein M. (1990). Le dictionnaire électronique des mots composés. *Langue française*, (87), 71-83.
Silberztein M. (2010). La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. *Langages*, (3), 221-241.
Silberztein M. (2016). Formalizing Natural Languages: the NooJ approach. Wiley Eds, Hoboken, NJ.
Silberztein M. (2018). Using linguistic resources to evaluate the quality of annotated corpora. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing* (pp. 2-11).
Taylor A., Marcus M. and Santorini, B. (2003). The Penn treebank: an overview. *Treebanks*, 5-22.
Volokh A. and Neumann G. (2011, June). Automatic detection and correction of errors in dependency treebanks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 346-350).

Validación de léxico en lengua vasca mediante análisis lexicométrico de un corpus paralelo multilingüe

Juan Abasolo, Naia Eguskiza, Aitor Iglesias

UPV/EHU – juan.abasolo@ehu.eus naia.egusquiza@ehu.eus
aitor.iglesias@ehu.eus

Abstract
This paper tests the first lexicon prepared for the use of the Reinert method in the Basque language, using a parallel multilingual corpus. For the construction of the aforementioned corpus, data were collected in five languages; on the one hand, in the Basque language (Euskera) and, on the other hand, in four languages with reference to the use of the Iramuteq software and the consequent uses of the corresponding lexicons (French, English, Spanish and Latin). The corpus consists of the Pauline letters of the New Testament, for which data were collected from vernacular translations of the Bible. The analysis of the letters was carried out in the same way in the different languages, in order to be able to compare the results in a valid way. The results show a high level of congruence between the different languages, although a number of differences can also be noted. It is believed that the different religious and cultural traditions in the translations of a single common text may also explain some of the differences in the results found.

Keywords: Basque language, dictionary, Reinert method, Iramuteq, multilingual corpus

Resumen
En este trabajo se pone a prueba el primer léxico preparado para el uso del método Reinert en lengua vasca, mediante el uso de un corpus multilingüe paralelo. Para la construcción del citado corpus se recogieron datos en cinco lenguas; por un lado, en idioma vasco (euskera) y por el otro, en cuatro lenguas con referenciada utilización del software Iramuteq y los consiguientes usos de los léxicos correspondientes (francés, inglés, castellano y latín). El corpus consta de las cartas paulinas del Nuevo Testamento, para lo que se recobraron los datos de traducciones vernáculas de la Biblia. El análisis de las cartas se realizó de idéntica manera en las diferentes lenguas, con el objetivo de poder comparar los resultados de forma válida. Los resultados muestran un alto nivel de congruencia entre las diferentes lenguas, aunque también se pueden señalar una serie de diferencias. Se estima que las diferentes tradiciones religiosas y culturales en las traducciones de un único texto común, también pueden ser explicativas de algunas de las diferencias entre los resultados encontrados

Palabras clave: euskera, léxico, método Reinert, Iramuteq, corpus multilingüe

1. Introducción

En este trabajo se pone a prueba el primer léxico preparado para el uso del método Reinert (Reinert, 1983) en lengua vasca, mediante el uso de un corpus multilingüe paralelo.

El análisis según el método propuesto por Reinert, hace años que se viene realizando en diferentes lenguas; en un comienzo con el software Alceste (Reinert, 1990) y, de hace unos años hacia acá, con la presentación del software libre Iramuteq (Ratinaud & Déjean, 2009) se ha extendido su uso a otros entornos, ambientes y lenguas.

Nos atañe este último punto, debido a que para cada nueva lengua que se desea agregar es necesaria la construcción de un léxico que permita procesar los textos lematizadamente (Ratinaud, 2021).

En este trabajo se da cuenta del análisis realizado a un corpus multilingüe compuesto por cinco módulos de las versiones de textos equivalentes escritos en otras tantas lenguas, con el objeto de señalar los elementos de similitud y de diferencia que produce cada uno de los análisis.

2. Método

En los próximos párrafos se presenta el detalle de la construcción del corpus, así como el procesamiento al que se expuso y las opciones que se tomaron en cuenta para la construcción del análisis según el método Reinert.

2.1. Corpus

El corpus lo componen cinco traducciones de las epístolas paulinas, tomadas de otras cinco versiones en cinco lenguas diferentes, tres de carácter altamente prepositivo (francés, inglés y castellano) y dos aglutinantes (latín y euskera).

Como se viene presentando, el objetivo final es la comparación del desempeño del léxico en lengua vasca con el que se constata en otras lenguas. Este objetivo condicionó la elección de los elementos del corpus multilingüe, restringido por el acceso a un corpus que incluyera el euskera; acto seguido, los módulos del corpus en otras lenguas tendrían que poder ser tratados de modo uniforme. Recalcamos el interés en incluir alguna otra lengua aglutinante, siendo el latín la única que cumple este requisito, entre las que se obtiene el léxico para ser utilizado en Iramuteq.

Las obras que componen el corpus son todas las epístolas paulinas del Nuevo Testamento, tanto las que tradicionalmente se atribuyen a San Pablo como las que hoy se presuponen póstumas.

Tabla 1. Resumen de los módulos del corpus paralelo multilingüe

Lengua	Tipología	Traducción	Año
Euskera	aglutinante	EAB	1994
Latín	aglutinante	Sacra Vulgata	382
Francés	prepositivo	Bible Segond	1910
Castellano	prepositivo	Dios Habla Hoy (latam)	1966
Ingles	prepositivo	Biblia del rey Jacobo	1611

Las características estilísticas diversas que se pueden intuir viendo la tabla 1, en la que se presentan los datos sobre cada módulo paralelo, condicionaron también el tratamiento posterior a la recogida de datos. Así, por un lado, se buscó producir un etiquetaje coherente, basado en las siguientes características de las unidades: la carta, el capítulo y el o los versículos parte de las unidades de análisis/contexto. Para esto realizó un pre-procesamiento automático con objeto de dividir el texto en unidades contextuales, según las características de la puntuación, y adjudicar las tres etiquetas antes mencionadas. La principal divergencia es que la versión latina, Sacra Vulgata, tiene tantas unidades como versículos, a la vez que carece de signos de puntuación más allá de la división en veresículos. Opuesta a esta versión están las otras cuatro, que en muchos casos reúnen más de un versículo en las unidades de sentido, entendidas desde la perspectiva de los cánones literarios establecidos en la puntuación.
Los diccionarios que se utilizaron fueron los que portaba consigo la instalación del software Iramuteq (Ratinaud, 2014) para todas las lenguas menos la lengua vasca. Para ésta última, se utilizó la versión de prueba anteriormente mencionada.

2.2. Tipo de análisis

El análisis está dividido en dos momentos. Un primer momento en el que se expone cada uno de los módulos del corpus a la Clasificación Jerárquica Descendente, referido ampliamente como CHD en la literatura y este artículo de ahora en adelante; y el segundo momento, en el que se contraponen estas clasificaciones en búsqueda de patrones de similitud o divergencia.
Hay un único análisis al que se expusieron los diferentes módulos del corpus. En cada uno se realizó un análisis de acuerdo al método Reinert, con el correspondiente léxico, construyendose así una CHD para cada módulo de análisis, siguiendo los siguientes parámetros:

- Clasificación simple sobre el texto
- Los tamaños de los segmentos de texto (RST) = 12 y 14

- Numero de clases al final de la fase uno = 25
- Numero máximo de formas analizadas = 3000
- Número mínimo de segmentos terminales de la fase 1 = 25
- Algoritmo para la descomposición: *irlba*

El resultado de esos análisis es el que se utilizó de base para un **segundo análisis** en busca de las relaciones que se pudieran evidenciar. Esta búsqueda se basó en el estudio de la organización de las etiquetas descriptoras, después de realizar las CHDs. Debido a que la oposición es entre cinco clasificaciones de textos en otras tantas lenguas, ni las formas activas ni las suplementarias pueden ser tomadas por comunes; de ahí la utilización del etiquetado común a todos los módulos del corpus.

El segundo análisis se realizó como análisis de clústeres, basado en la distancia entre cada clase de las CHD. Como primer paso se computó una matriz de distancias entre los clústeres, calculada con el algoritmo de Jaccard, sobre una matriz de presencia/ausencia de las etiquetas en esas clases. En el siguiente paso se calcularon los clústeres sobre esa matriz, utilizando el algoritmo de Ward aplicado en R (Murtagh & Legendre, 2014), según el paquete base de R; esta es una clasificación jerárquica ascendente. Para finalizar, se determina el mismo número de clústeres que los que el método Reinert reveló para el corpus del léxico objeto de estudio, en lengua vasca.

El primer análisis dejó entrever, como se puede ver en la tabla 2, que las propias características tipológicas de las lenguas, además de las de los léxicos que se usaron con estos, tienen reflejo en las características cuantitativas que devuelve el CHD. Esa característica se refleja claramente en el número total de formas, más altas en latín y en euskera, lenguas aglutinantes, que en francés, inglés y castellano. A su vez, también se ve en las ocurrencias totales, que son menores en las lenguas aglutinantes que en las que no lo son.

Tabla 2. Descripcion de los cinco análisis

léxico	textos	seg.	form.	ocur.	lemas	f. activas	f. sup	f.act.>3	f. por. seg.	nº clases	seg. clasificados
fr	2336	2438	4437	46342	2699	2427	272	1009	19.00	7	2278
en	2967	2972	2881	48837	2126	1795	331	880	16.43	12	2825
es	2663	2669	4759	45740	2492	2368	121	996	17.14	9	2530
la	2021	2021	5984	30209	2846	2735	108	1064	14.95	12	1887
eu	2669	2669	6544	32143	3759	3453	178	1124	12.04	11	2515

Un análisis exploratorio de los resultados permite corroborar una correlación significativa entre la cantidad de segmentos del módulo (*seg*) y el total de

segmentos clasificados (*seg. clasificados*) en el CHD. La relación más baja se ve en el módulo en latín, con 9,3 textos de cada 10 y la más alta en el módulo en inglés con una proporción de 9,5 sobre 10. Se calculó la correlación de Pearson entre los datos obteniendo el resultado r=0.999, con un p-valor < 0.001.
Entre el total de las formas activas detectadas en el análisis, Iramuteq también ofrece información acerca de cuántas de estas se repiten tres o más veces en el corpus analizado. Aquí también se encontró una relación entre el número de formas detectadas en cada módulo y la cantidad total que se repiten tres o más veces, pero es de destacar que las CHD realizadas en francés o inglés detectaron que casi la mitad de las formas activas se repetían tres o más veces. En oposición a este resultado encontramos que sólo un tercio de las formas detectadas como activas en la clasificación del módulo vasco se repite tres o más veces.
Con respecto al número de clases (*nº de clases*) en las que se clasificó cada módulo del corpus, no se encontró relación significativa con ninguna de las otras variables descriptivas.

2.3. Estudio de las clasificaciones

La metodología que se utilizó para enfrentarnos al elemento de análisis, demostró un grado de coherencia interléxica similar.
El número de clases encontradas entre las cinco clasificaciones de los otros tantos módulos del corpus, devolvió cincuenta clases en total, detalladas en la tabla 2. Estas clases se reagruparon según su similitud siguiendo el método de agrupamiento de Ward y se retuvieron once clústeres, la cantidad de clústeres conseguida en el CHD en la versión vasca.
En el dendrograma que se presenta en la figura 1 se remarcan las clases obtenidas en la clasificación de la versión vasca utilizando el color verde. Se puede ver que el primer agrupamiento que se genera en la clasificación Ward es el que une a las quintas clases de vasco *eu5* y castellana *es5*, que luego de juntarse a las primeras francesa *fr1* e inglesa *en1*, son el último clúster en unirse al tronco común.

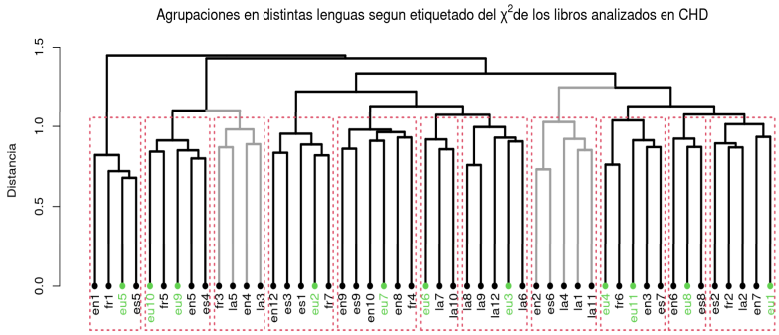


Figura 1. Clasificacion jerarquica ascendente de las CHD en las cinco lenguas. 11 clusterés categoricos resaltados

Si bien no era objeto de este estudio, merece ser recalcada la evidencia de que las clases de la CHD del texto latino, el de factura más temprana de los cinco, son las únicas que en las agrupaciones ascendentes mayormente se unen entre sí.

Caso del primer clúster

El caso del primer clúster parece ser el más ilustrativo, ya que en el propio dendrograma de la figura 1 se constata que es, desde la perspectiva ascendente del algoritmo de Ward, el primer grupo que se conforma a la vez que es el último que se une al total. Este primer clúster está formado por una única clase de las conformadas en los CHD de los módulos vasco, castellano, francés e inglés.

Figura 2. Nubes de palabras de los elementos del primer clúster

El análisis de la distribución de las etiquetas revela que la *Primera carta a los corintios* es común en las cuatro primeras clases, específicamente, los capítulos 8, 10 y 11 y los versículos 18, 20, 21 del capítulo 10 y los versículos 22 y 25 del capítulo 11. También son comunes a las cuatro clases, el versículo 8 del 3º capítulo de la Carta a los tesalonicenses y el versículo 20 del capítulo 12 de la Carta a los romanos. Las siguientes etiquetas no son comunes a las cuatro clases sino a tres o menos. En la figura 2 se presentan las cuatro nubes de palabras de las formas activas de los elementos del clúster.

3. Conclusiones

En este trabajo se presentó una comparación entre los análisis de cinco textos equivalentes en cinco lenguas, con el objeto de testear la coherencia de los resultados de la utilización del léxico vasco.

Se puede apreciar que el comportamiento de los análisis hechos con los ya probados léxicos en francés, inglés y castellano, no difiere significativamente del comportamiento obtenido con el léxico puesto a prueba. Sí se puede comprobar, igualmente, que el análisis del texto latino arroja resultados diversos. Poniendo el foco en el léxico evaluado, es de relevar que la relación entre formas activas detectadas y formas activas con ocurrencia mayor a tres veces es más baja en el léxico estudiado que en el resto. Eso señala la hipótesis de que es

posible una multiplicidad innecesaria de lemas en el léxico utilizado.
Con el mismo espíritu, remarcaremos un límite detectado a posteriori en la elección y procesamiento del corpus. Las diferentes características inter módulos, debido a la falta de unificación de criterios estilísticos en las diferentes versiones de la Biblia podría ser una fuente de ruido en los datos. Eso deberá ser tomado en cuenta para replicar el estudio.
Por último, es de señalar que con idénticos parámetros no se encontró relación entre el número de clases resultantes y ninguna de las variables estudiadas, esta característica no deja de merecer ser objeto de estudio en futuros análisis de corpus paralelos.
A la luz de este trabajo, se estima prudente promover el uso del análisis según el método Reinert aplicando el software Iramuteq de textos y cuestionarios en lengua vasca con el léxico objeto de estudio.

Referencias

Murtagh F. and Legendre P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274–295. https://doi.org/10/f6p65wX

Ratinaud P. and Déjean S. (2009). IRaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. *Modélisation appliquée aux sciences humaines et sociales MASHS*, 8–9.

Ratinaud P. (2014). IRaMuTeQ: *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* [Linux]. http://www.iramuteq.org

Ratinaud P. (2021, marzo 13). Re: *[Iramuteq-users] Générer le lexique pour une outre langue*. https://sourceforge.net/p/iramuteq/mailman/message/37239005/

Reinert A. (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.

Reinert M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 26(1), 24–54. https://doi.org/10/cbhfwp

Construcción de un léxico en lengua vasca para su utilización con el software IRaMuTeQ

Juan Abasolo, Naia Eguskiza, Aitor Iglesias

UPV/EHU – juan.abasolo@ehu.eus naia.egusquiza@ehu.eus
aitor.iglesias@ehu.eus

Abstract
The adaptation of the Reinert Method, using the Iramuteq software, in each new language requires an adapted dictionary for its use. This paper presents the challenges and choices made in the construction of a standard lexicon for the use of Iramuteq with texts written in Basque (Euskera). Unlike prepositional languages of Indo-European origin in which it has a proven usage, Basque is of a highly agglutinative typology. The recovery of lemmas in prepositional languages requires a much smaller number of tokens (forms) than in the case in question. Three main sources, corresponding to different linguistic moulds, were used to build up the Basque dictionary. The total corpus for the creation of the lexicon was more than 1,150,000 sentences and was analysed using Universal Dependences and other applications of the computer language R. This made it possible to generate a primordial lexicon of more than half a million forms. Today, the lexicon still needs to be modified on the basis of the general and common criteria of users in the Basque language, for which a work proposal is presented.

Keywords: basque language, dicctionary, Reinert method, Iramuteq

Resumen
La adaptación del Método Reinert, utilizando el software Iramuteq, en cada nueva lengua necesita de un diccionario adaptado para su uso. En este trabajo se presentan los desafíos y las opciones tomadas en la construcción de un léxico estándar para el uso de Iramuteq con textos escritos en lengua vasca (euskera). A diferencia de las lenguas prepositivas de origen indoeuropeo en las que tiene probado uso, el vasco es de tipología altamente aglutinante. La recuperación de los lemas en lenguas prepositivas necesita de una cantidad de tokens (formas) mucho menor que en el caso del euskera. Para la constitución del diccionario en lengua vasca se utilizaron tres fuentes principales, que responden a diferentes moldes lingüísticos. El total del corpus para la creación del léxico fue de más de 1.150.000 oraciones y fue analizado mediante Universal Dependences y otras aplicaciones del lenguaje informático R. Eso permitió generar un léxico primigenio de más de medio millón de formas. Hoy en día el léxico todavía necesita modificaciones en base a los criterios generales y comunes de los usuarios en lengua vasca, para lo que se presenta una propuesta de trabajo.

Palabras clave: euskera, léxico, Metodo Reinert, Iramuteq

1. Introducción

El objetivo de este trabajo es construir un léxico operativo en lengua vasca, para utilizar el software IRaMuTeQ (Ratinaud, 2014) sobre textos escritos en esa lengua, según los cánones y usos actuales dentro del registro formal.

Ya existe más de un estudio publicado en vasco en el que se utilizó el método Reinert (Navarro e Idoiaga, 2021; Larruzea-Urkixo et al., 2019), así como una serie de trabajos basados en datos recogidos en lengua vasca que no pueden llevarse a cabo sin un exhaustivo trabajo de pre-procesamiento de los datos.

Para la realización de los análisis multidimensionales de textos o cuestionarios en una determinada lengua con todo el potencial de Iramuteq, se necesita disponer de un repertorio de las formas que se esperan encontrar en los textos a analizar, con sus correspondientes valores semánticos (lemas), así como la categoría gramatical que se espera que se le atribuya a ese valor semántico. Esta relación tiene también que ser univoca y excluyente, ya que ninguna forma del repertorio en cuestión puede pertenecer a más de un lema ni ningún lema puede pertenecer a más de una categoría gramatical.

A fecha de hoy, la instalación de la versión 0.7 alpha 3 para Linux del software en cuestión proporciona dos accesos diferentes a los diccionarios de estas lenguas: francés, inglés, italiano, portugués, castellano, alemán, sueco, griego y gallego (los últimos cuatro de manera experimental). Además de los léxicos de esas lenguas se puede acceder a los léxicos en rumano, latín, y a otra versión del léxico en griego.

A excepción del caso de la lengua latina, el resto de lenguas mencionadas son lenguas prepositivas, lo que implica que la cantidad de formas por lema sea menor que en las lenguas aglutinantes, ya que la variabilidad morfológica es menor. Valga de ejemplo recordar que *rosa* y *rosas* en castellano, del lema *rosa*, en latín se multiplican en *rosa, rosae, rosam, rosas, rosarum*, y *rosis*.

Seguidamente presentamos algunas características relevantes de la lengua vasca, vinculadas tanto con temas intrínsecamente lingüísticos, como también sociolingüísticos.

1.1 El vasco y la tipología aglutinante

La lengua vasca alcanzó una norma estandarizada en el año 1968. Ese estándar construido, llamado “euskera unificado”, es el que vehiculiza la educación. La lengua vasca, es de características altamente aglutinante, lo que implica que, a diferencia de las lenguas vecinas, gascón, francés o castellano, no utilizan preposiciones, sino sufijación.

A la hora de determinar las formas que puede tomar un lema en el lenguaje escrito, su número en euskera es mucho más alto que en otras lenguas. Así, en castellano basta con dos formas para ocupar todas o casi todas las ocurrencias

escritas del lema *casa*+, en vasco es necesario un número mucho más amplio solo para los usos más extendidos; *etxe, etxea, etxeak, etxeko, etxetik, etxera, etxean, etxerako, etxetan, etxeetan… [casa, (la) casa, (las) casas, de casa, desde casa, a casa, en casa, para casa, en casa, en casas…]*.

1.2 La geografía, el léxico y las morfologías nominal y verbal

El euskera tiene, siguiendo a Aurrekoetxea et al. (2019) tres dialectos o siete subdialectos. Cada una de estas variedades tiene su propia variedad léxica, característica que se refleja en mayor cantidad de lemas y formas en algunos campos semánticos; *rápido*, que puede ser mencionado como *arin, agudo, azkar, fite* o *zalu*. A esa multiplicación léxica se le debe sumar la morfología nominal, en la que el mismo lema, *ama* [mamá] puede tomar tres formas geolingüísticamente marcadas, sin diferencias de sentido, *amarentzat, amarentzako, amarendako* [para mamá], siempre refiriéndose al único lema mencionado.

2. Proceso de construcción del léxico

El léxico que se utiliza en la aplicación Iramuteq debe ser un archivo de texto plano con al menos tres columnas separadas por tabulación. La primera columna debe tener las formas o tokens que se espera que aparezcan en el corpus a analizar. La segunda columna, los lemas a los que se refieren esos tokens. La tercera columna debe tener las categorías gramaticales, etiquetadas en francés. Optativamente, también puede tener más columnas con información relevante acerca del léxico, que no participarían en el análisis.

2.1. Herramienta

Para la construcción de la mayoría de los léxicos, se propone (Ratinaud, 2021) la utilización del software TreeTagger[1]. Ese software que permite etiquetar un corpus dado al día de hoy no dispone de analizador para el euskera.

Dentro del proyecto Universal Dependencies[2] (McDonald et al., 2013) existe un modelado para el euskera (Aranzabe et al., 2015), que es, a su vez una adaptación del original AbarrHItz de Ilarraza et al. (2004). Esta herramienta permite anotar, mediante algunos paquetes informáticos, como Udpipe (Wijffels, 2020), corpus de gran tamaño, con un nivel de precisión del 71,27 %.

[1] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[2] https://universaldependencies.org/

2.2. Corpus

El corpus está compuesto por tres subcorpus, pretendiendo de este modo abarcar gran parte de la variabilidad de usos que se da en la lengua vasca. El primer subcorpus es el tomado del proyecto CommonVoice de Mozilla Fundation, compuesto por 61.567 frases seleccionadas para ser leídas a viva voz y evaluadas. El segundo subcorpus está compuesto por el total de los artículos que se publicaron durante el año 2020 en la publicación on-line Bizkaie[3]. Esa publicación es la única que hoy en día publica casi diariamente en la variedad estándar del dialecto occidental. De ese corpus se obtuvieron un total de 11.379 frases. El tercer subcorpus está compuesto por la totalidad de los artículos publicados en línea por la revista de información generalista Argia[4] (1989-2021) incluye más de un millón de frases y abarca varios registros del euskera estándar.

2.3. Análisis y construcción del léxico

La construcción se planteó de manera modular, siendo cada uno de los corpus tratados entendido como un módulo independiente, que a la hora de terminar la construcción deberían ser juntados en un único elemento coherente.

Para enfrentar ese proceso, cada uno de los corpus fue dividido en frases analizables sintácticamente mediante un guión escrito en R. Con ese objeto los criterios de tratamiento de los datos fueron los determinados por la puntuación. De ese primer tratamiento de los datos se tomaron todas las formas resultantes factibles de ser parte del léxico final, así como los valores lemáticos y las categorías gramaticales adjudicadas a estas. En esta tarea se descartaron, además de toda la información contextual, las informaciones de puntuación y la información numeral.

Ese proceso produjo tres protoléxicos que carecían de la coherencia y univocidad que se exige en el uso de Iramuteq: Cada forma adjudicada a un solo lema y cada lema único adjudicado a una única categoría gramatical.

El criterio a observar en este proceso de selección y limpieza, fue el cuantitativo. En función de un uso mayor de un lema sobre otro, para la misma forma, la forma se adjudica al lema reportado con mayor frecuencia; de idéntico modo, cuando un lema es codificado en más de una categoría gramatical, la que se le adjudica es la de mayor frecuencia. Este primer procesamiento llevó a una reducción desde las 13.065.741 coincidencias de formas, lemas y categoría gramatical, a unas 627.479. Finalmente, el último paso fue el de traducir los apócopes ingleses adjudicades por el paquete udpipe a los franceses que necesita Iramuteq para funcionar.

[3] https://www.bizkaie.biz
[4] https://www.argia.eus

Se optó por reducir las etiquetas de categoría gramatical, a las que se utilizan en los léxicos inglés o portugués, de probado uso. Aunque, debido a las características aglutinantes de la lengua, se consideró pertinente conservar las múltiples formas de los nombres propios de uso corriente. Para eso, la etiqueta *PRONOM* de udpipe se transformó en *pro_per*, identificada en el léxico francés como pronombre personal y que no es activa por defecto en los análisis de Iramuteq.

3. Resultado

En este trabajo el léxico que estamos presentando tiene un total de 526.046 formas diferentes tomadas de los tres entornos ya mencionados en los que se utiliza la lengua vasca escrita. Esas más de medio millón de formas se muestran en correspondencia con 221.255 lemas, de los que más de la mitad, 133.205, se corresponden con nombres propios.

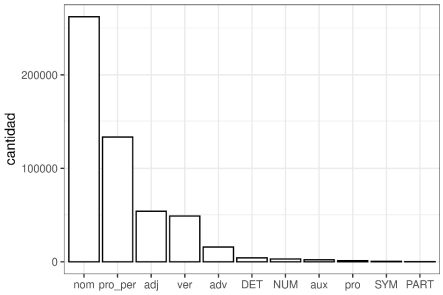


Figura 1. Frecuencias de formas según categorización gramatical

Como se puede observar en la figura 1, se conservaron algunas categorías gramaticales asignadas en el análisis sintáctico, que no se incluyen en el elenco de categorías analizables con el software Iramuteq, las etiquetadas como DET (determinante), NUM (numérico), SYM (símbolo) y PART (partícula). Se prescindió de traducirlas, después de haber detectado que gran parte de las formas responden a otros lenguajes, entremezclados con los propios de la lengua. Cabe destacar que la categoría PART (94 formas) se refiere a algunas partículas que pueden o no ser adverbios, tales como *bai* [sí], *ez* [nó], *omen* [partícula que denota incerteza].

3.1. Comparación

Como prueba de la posible efectividad del léxico en el análisis de textos escritos en lengua vasca, Abasolo y Eguskiza (2022) presentaron una clasificación jerárquica descendente de las autodescripciones escritas por los docentes de los grados de educación de la Universidad del País Vasco (UPV/EHU). Resultó

imposible replicar el análisis sin lematización, por eso la comparación está basada en la descripción de las estadísticas.

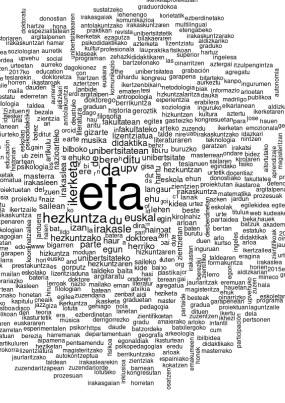


Figura 2. Nube de palabras del mismo corpus sin lematizar.

Figura 3. Nube de palabras del mismo corpus lematizado.

En las nubes de palabras de las figuras 2 y 3 se aprecia una gran diferencia en las palabras centrales. Por un lado, en el texto sin lematizar (figura 2) se observan la centralidad de formas de la conjunción *eta* [y] y de los verbos auxiliares (*da* o *du* [3ª pers sing]) y por otro, se disgrega *hezkuntza* [educación]. En el análisis lematizado (figura 3) *hezkuntza* aparece 487 veces y en el tratamiento sin lematizar se cuentan 12 formas (h*ezkuntz, hezkuntza, hezkuntzako, hezkuntzan, hezkuntzara, hezkuntzarako, hezkuntzarekin, hezkuntzaren, hezkuntzari, hezkuntzak, hezkuntzatik, hezkuntzazko*), con entre una y 263 ocurrencias. El análisis lematizado detectó un total de 3.624 elementos, 2.225 de una única incidencia, y el análisis sin lematizar 5.067 elementos, de los que 3.235 aparecieron una única vez.

3.2. Límites

El análisis pormenorizado del léxico pone a la luz que el 71 % de precisión en el análisis de la lengua vasca al que alcanza el parser utilizado, también conlleva un 29% de imprecisiones. Así, la revisión manual puede identificar formas adjudicadas a lemas errados, infravalorando otros. Sirve de ejemplo señalar el inexistente verbo **ekartu* en el que se clasificaron las formas *ekartzekotan* y *ekartzerakoan* que debieran ser clasificadas dentro del lema *ekarri* [traer], al que se le adjudicaron correctamente cuarenta y tres formas.

No se realizó ningún léxico auxiliar de expresiones para la detección correcta de las formas perifrásticas, que en euskera es muy extendida. Se prevé que se van a sobredimencionar algunos verbos, tales como *egin* [hacer] o *hartu* [tomar/agarrar] sobre los que se construyen multitud de perífrasis verbales, tales

como *hitz egin* [hablar], *igeri egin* [nadar], *lo egin* [dormir] o *loak hartu* [dormirse], *parte hartu* [participar] etc.

3.3. Proceso de adecuación

Prevemos que esta versión del léxico en lengua vasca que se presenta va a necesitar sucesivos ajustes tanto para corregir las dificultades arriba mencionadas, como también para abarcar usos corrientes de la lengua que no hayan sido retenidos entre los datos aquí presentados.

Con ese objeto, se propone una herramienta para la corrección colectiva en la que se presenta un formulario de respuesta rápido para identificar los elementos a corregir o agregar y los scripts para que, a partir de esas respuestas, se genere la nueva versión del diccionario, que será publica y descargable en el repositorio de acceso abierto Open Science Framework.

En este artículo se ha presentado y dado cuenta de la construcción del primer léxico operativo en lengua vasca para utilizar el software para análisis textual y de cuestionarios Iramuteq. Este es el primero que se construye en una lengua aglutinante viva intentando responder a retos propios de ésta, aún debe ser puesto a prueba y mejorado, para lo que se presentan las herramientas y el proceso a seguir.

Referencias

Abasolo, J., y Eguskiza, N. (2022, abril 6). I*ramuteqerako euskarazko lexikoa V0.1, aurkezpena* [Comunicación]. Goi-Mailako Hezkuntzan Berrikuntza Naziorteko Kongresua, Leioa.
http://eudia-ehu.github.io/aurkezpenak/2204GoiMailakoHezkuntza/lexikoa-iramuteq/

Aranzabe, M. J., Atutxa, A., Bengoetxea, K., Diaz de Ilarraza, A., Goenaga, I., Gojenola, K., & Uria, L. (2015). Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In M. Dickinsons, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Arg.), *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories* (TLT14) (pp. 233–241). Institute of Computer Science of the Polish Academy of Sciences.

Aurrekoetxea, G., Gaminde, I., Ormaetxea, J. L., & Videgain, X. (2019). *Euskalkien Sailkapen Berria*. UPV/EHU, Argitalpen Zerbitzua.

Epitafio de Antestius Sesenco. (s. f.). https://www.wikidata.org/wiki/Q97345601

Ilarraza, A. D. D., Garmendia, A., & Oronoz, M. (2004). Abar-hitz: An annotation tool for the basque dependency treebank. *In Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Lafitte, P. (1921). Les vieux textes basques. Les «Glosas Emilianensis». *Gure herria : bulletin de l'Association «Gure Herria»*., XIII, 11–12.

Larruzea-Urkixo, N., Cardeñoso, M. O., & Idoiaga, N. (2019). Emozioak unibertsitateko ikasketa prozesuen ardatz: Lehen Hezkuntzako Graduaren kasua. *Tantak. EHUko Hezkuntza Aldizkaria.*, 31(1). https://doi.org/10.1387/tantak.20578

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev,

K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Castelló, N. B., & Lee, J. (2013, enero 1). *Universal Dependency Annotation for Multilingual Parsing*. ACL (2). https://openreview.net/forum?id=S1-bzTxuWH

Musée Saint-Raymond, Ra 385. (27a.C.). Musée Saint-Raymond. https://commons.wikimedia.org/wiki/File:MSR-Ra385-FLP.tif

Navarro, G., e Idoiaga, N. (2021). Bertso-eskolak, nerabezaroan hezteko espazio gisa. *Uztaro: giza eta gizarte-zientzien aldizkaria, Uztaro 117*, 75–90.

Ratinaud, P. (2014). IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Linux, Windows, Apple]. http://www.iramuteq.org

Ratinaud, P. (2021, marzo 13). Re: [Iramuteq-users] Générer le lexique pour une outre langue. https://sourceforge.net/p/iramuteq/mailman/message/37239005/

Wijffels, J. (2020). *udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the «UDPipe» «NLP» toolkit* [Manual]. https://CRAN.R-project.org/package=udpipe

A hybrid approach to opinion mining in Facebook's disinformative echo chambers

Suania Acampa

University of Naples Federico II – suania.acampa@unina.it

Abstract
In most studies on social media disinformative echo chambers, users are considered not as different persons who respond to complex social behaviors, but as a homogeneous tribe with a single thought. The aim of research is to understand the cultural and communicative processes that emerge in the disinformative environments of social media analyzing users to a greater degree of depth: investigating how and in what direction their opinions are radicalized through the reconstruction of their narratives. For this purpose, it was developed a hybrid approach to opinion mining able to combine the researcher's full control over the classification process, maximizing its capabilities on a large scale. Thanks to a hybrid approach, it was possible to provide disinformation with a further interpretative aspect, framing it in terms of the ideological split of users.

Keywords: disinformation, echo chambers, opinion mining, text analysis, sentiment analysis.

1. Introduction

The image of echo chambers aims to convey the idea of the people who use social media platforms are exposed largely (or exclusively) to one type of pro-aptitude content. On echo chambers, we know from the literature that user communities polarize around specific topics (Quattrociocchi & Vicini, 2016). There is a very productive line of studies - fully inserted in the field of computational social sciences (Lazer et al., 2009) - which aims to understand the social dynamics that emerge among users who consume misinformation on social media platforms. This empirical line analyzes social dynamics on the platforms using Social Network Analysis; it is characterized by an exclusively data-driven approach whose purpose to exploit the large amount of data coming from social media to build mathematical models and computational tools precise to the point of understanding, anticipating and hypothetically controlling the behavior of users in these echo chambers (Quattrociocchi, 2021). This approach boasts of having empirically detected the very existence of echo chambers and demonstrated the role of user confirmation bias in information selection (Mocanu et. al., 2015; Bessi et. al., 2015; Bessi et. al., 2014). What we don't know is if these users inside their echo chambers all behave the same way or if there is some element that can differentiate them, because an exclusively data-driven approach considers

different individuals and complex social behaviors like a single subject user with undifferentiated behavior. There is little empirical research attempting to investigate how and in what direction user opinions are radicalized. This is what we have tried to do in this study. The aims of the work are: 1) Investigate the communicative processes of Italian users who consume disinformative content. 2) Reconstruct users' narratives and worldviews that emerge during their interaction practices with disinformative content. 3) Identify the factors that guide users in their experience inside disinformative echo chambers.

2. Methodology

The analysis focuses on the Facebook platform. Ten disinformative Italian pages were identified, and the data was collected via API from January 2016 (known as the year of disinformation) to April 2020. The population reached is 18,767 posts and over one and a half million comments (precisely 1,549,872). The sample was constructed on 10% of the population and extracted through a type of proportional stratified sampling for a total of 1,877 posts with their respective 112,000 comments. The analysis was conducted on two levels: the first level of analysis focused attention on the transmission of the message (stimulus) while the second level of analysis, which this paper deals with, focuses on the involvement of users as a response to the stimulus message. In the first level, the posts were labeled based on the disinformation risk (high, medium or low). In the second level of investigation (linked to the first by an anchor variable) users' opinions and the polarization of their narratives were investigated. For this aim, two typical approaches of opinion mining were used: the supervised machine learning approach based on the probabilistic classifier Naive Bayes, historically one of the most used algorithms in the automatic classification of text; an unsupervised approach using the LIWIC dictionary to label texts on the words associated with psychologically and emotionally significant categories, in the end was adopted a hybrid approach, developed to cope with the inability of the classifier chosen to label classes with reference to particular categories.

2.1. Naïve Bayes classifier

To training set construction to be submitted to the Bayesian classifier, a sub-sample of comments composed of 5,604 comments (equal to 5% of the comments of the sample) which were manually labeled based on the sentiment expressed (positive, negative, neutral) and based on the political orientation expressed (right, left, center, political orientation rejected or not expressed). The classification capacity of the algorithm has been verified through the synthetic indicators (based on the confusion matrix) which contain information on the predictive capacity of the classifier. The Bayesian classifier returns an 82% accuracy in forecasting the sentiment variable, with a 95% probability that the

true accuracy for this model is in the confidence interval of 0.80 to 0.83 and an accuracy of over 80% on the three classes, which refers to the ability of the classifier not to label an instance with one label when it belongs to another. There is a certain difficulty in predicting the neutral and the positive, this is because the training set randomly extracted is composed of only 3% of the neutral class and 20% of the positive class against 77% of the negative class. The accuracy of the classifier drastically decreases in the prediction of political orientation (53% with a NIR of 72%). Despite the accuracy of 84.8% on the right political orientation, it has difficulty in predicting the other orientations (34.4% not expressed; 10.5 refuse placement; 0.66% on the left). The difficulty of the classifier may be due to the composition of the training set but also to the political scenario that has characterized Italy in recent years which has brought together very distant political positions. For this reason, a hybrid road was developed with an unprecedented combination of a supervised approach to machine learning and an unsupervised dictionary-based approach.

2.2. A hybrid approach to opinion mining

The hybrid approach exploits convergence to overcome the limits of one approach with the advantages of the other and vice versa (Amaturo & Punziano, 2016). Thus, starting from the labeled training set, a vocabulary of political orientation was built thanks to which it was possible to classify the comments by extracting their characteristics. However, the extraction of the characteristics is not so immediate because if on the one hand they must discriminate and describe the original text as much as possible, on the other hand they must also reduce the large size of the source data and avoid redundancies. For this reason, the function considered most suitable was the Frequency-Frequency Term Inverse Document- This feature is widely used in the information retrieval field to measure the importance of a word to a document in relation to a collection of documents. The difficulty with the selection or weighting of terms lies in establishing a good balance between popularity and specificity (Aizawa, 2003). The TF-IDF statistic appears to be the right compromise.

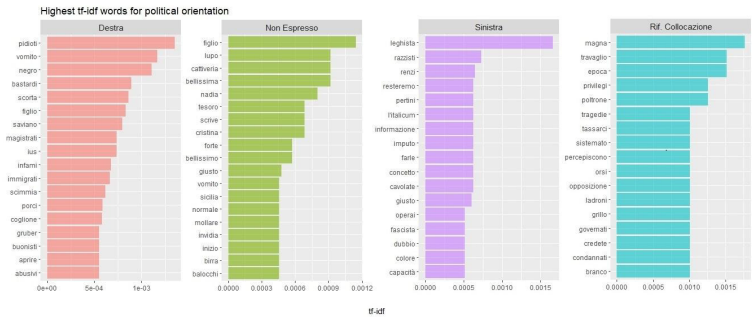


Figure 1. Words with the highest TF-IDF value based on political orientation

The graph shows only the first 15 entries with the highest value, but it is evident how the lexicon used for the different political orientations differs and how little it lends itself to ambiguity. The accuracy and the statistics of the model show that the results improved using this hybrid approach. (78% accuracy with a NIR of 48%). The sample's comments were thus labeled according to the sentiment expressed (through Naive Bayes), according to the emotions expressed as anger, anxiety, etc. (through the LIWIC lexicon) and according to political orientation (through the hybrid approach shown).

3. Analysis and Results

A factor and classification analysis (Gherghi & Lauro, 2004) were performed. The analysis of multiple correspondences highlighted two latent dimensions that are characterized by the opposition between high and low emotionality, the first factor (horizontal axis) and high and low conviction (second factor, vertical axis). The cluster analysis brings out the differences between the different users: right-wing political orientation (12%), left-wing political orientation (5%), political orientation refused (5%), political orientation not expressed (78%).

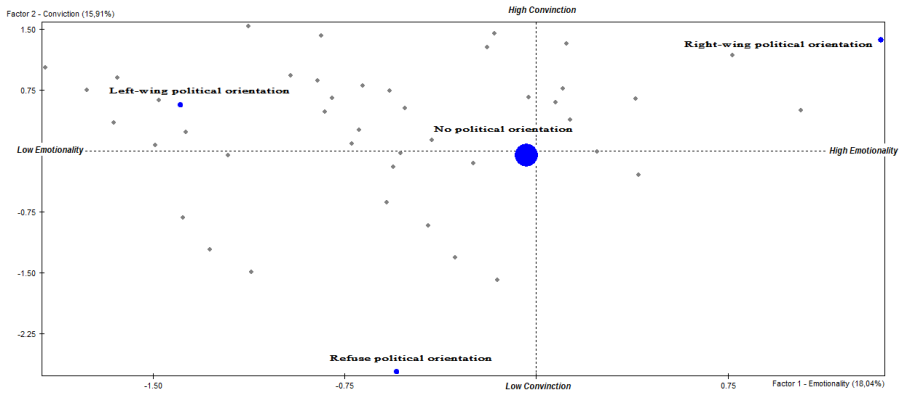


Figure 2. Users clusters on a factor graph deriving from ACM (axes 1-2)

Along the first axis, the high emotionality contained in user messages corresponds to the interaction with high-risk disinformation content that negatively polarize sentiment and generate very high user engagement (high engagement). The conviction expressed towards the opinions expressed by these users is also high. This negative communication polarized and convinced characterizes the political position of the right. The correlated feelings are anger and anxiety, the two emotions most frequently associated with populism and with major episodes of conspiratorial thinking (Grzesiak-Feldman, 2013). The high

conviction of right-wing users makes a comparison with different positions difficult. Their interaction makes these contents viral, keeping alive a very open debate on very delicate topics (extracted in the first level of analysis) such as immigration, Islam, LGBT rights, media, climate, etc. On the opposite side of the factorial plan, however, the low emotionality expressed in the users' comments corresponds to the interaction with content with a lower risk of disinformation that generate medium-low engagement. Here, left-oriented users have positive or neutral feelings. In the end, along the second axis, low conviction and low emotionality characterize users who reject political positions. The cluster analysis brings out the differences between the different users who appear as many small tribes who express different convictions and emotions, discuss different topics with different languages and expose themselves in different ways to the risk of disinformation.

3.1. The users' narrative

To understand on which issues the users' positions are polarized, the narratives of the different political orientations have been reconstructed with a Markov chain: a Text Mining tool that generates a semantic oriented discussion network.

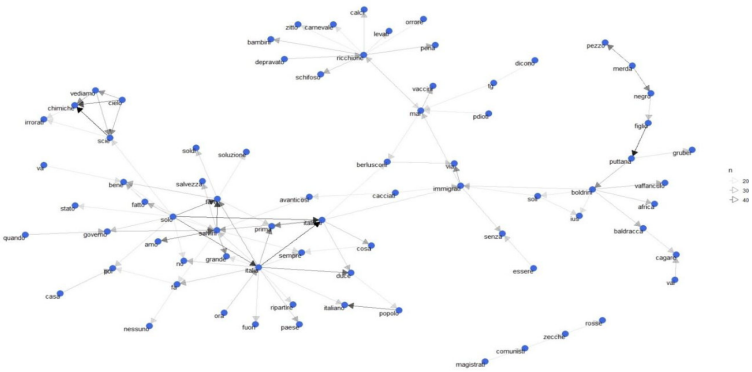


Figure 3. Markov network users with right-wing political positions (frequency ≥ 20)

The networks show that the right-wing users' narrative is characterized by strong radicalization and frequent references to hatred. It can be resumed in six elements: nationalism, conspiracy, homophobia, racism, misogyny, political opposition.

The left-wing users' narrative appears moderate, not very radicalized, and rather positive. There is a strong reference to the Italian historical left's figures (Berlinguer, Pertini) linked by their traditional battles (pensions, poverty, working class). The political opposition is above all towards Salvini and his party (Lega). The most interesting thing that emerges from this narrative is the

awareness of disinformation and the theme of fake news linked to Salvini.
The narrative of those who reject a political position appears less polarized than the first but more polarized than the previous one. The themes are those dear to the Movimento 5 Stelle: the "poltrone" (seats) and the salaries of politicians from which the anti-European and anti-elitist sentiment. Two souls also emerge: in favor of President Conte and against. The interesting thing is the presence (once again) of the theme of fake news linked to Salvini.
Users who do not express any political position have non-radicalized opinions on generic issues (such as TV, gossip, entertainment).

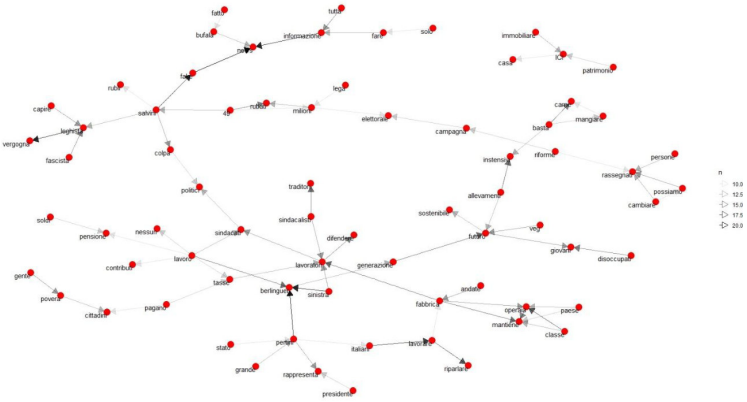


Figure 4. Markov network users with left-wing political positions (frequency ≥ 10)

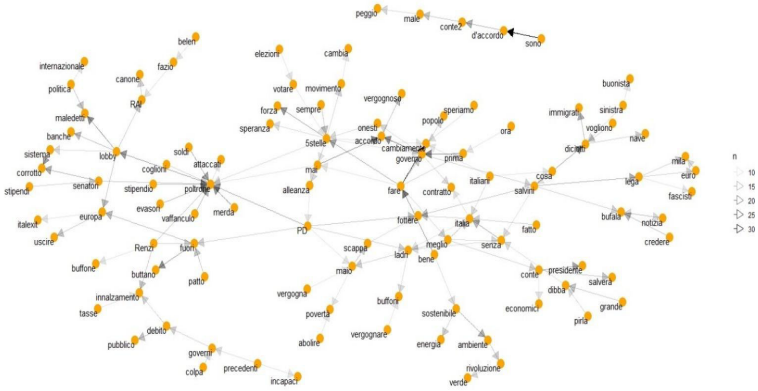


Figure 5. Markov network users who reject political positions (frequency ≥ 10)

4. Conclusion

In the light of what has emerged, it is possible to provide disinformation with another interpretative aspect: it is reasonably associated with populist support, and we can frame like an ideological split. Becomes clear how much political positions - with varying degrees of conviction and emotion - polarize and differentiate the response to the users' disinformative content in the echo chambers, thus becoming a guiding mechanism in the complex disinformative universe. The work certainly has limits, but the use of hybrid approach has empirically point out how the political positions differentiates users inside disinformative echo chambers and become their guiding mechanism in the complex disinformative universe.

References

Aizawa A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1), 45-65.

Amaturo E. and Punziano G. (2016). I *Mixed Methods nella ricerca sociale*. Roma: Carocci.

Bessi A., Caldarelli G., Del Vicario M., Scala A. and Quattrociocchi W. (2014), Social determinants of content selection in the age of (mis) information. *International Conference on Social Informatics* (pp. 259-268). Springer, Cham.

Bessi A., Coletto M., Davidescu G. A., Scala A., Caldarelli G. and Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2).

Gherghi M. and Lauro C. (2004). *Appunti di analisi dei dati multidimensionali. Metodologia ed Esempi*. Napoli: RCE Multimedia.

Grzesiak-Feldman M. (2013), The Effect of High-Anxiety Situations on Conspiracy Thinking. In *Current Psychology*, 32(1), pp. 100–118.

Lazer D., Pentland A. S., Adamic L., Aral S., Barabasi A. L., Brewer D., ... and Van Alstyne M. (2009). Life in the network: the coming age of Computational Social Science. *Science* (New York, NY), 323(5915), 721.

Mocanu D., Rossi L., Zhang Q., Karsai M. and Quattrociocchi W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior*, 51, 1198-1204.

Quattrociocchi W. (2021), L'era della (dis)informazione. *I Quaderni de le Scienze*, nr. 14, pp. 3-9

Quattrociocchi W. and Vicini A. (2016). *Misinformation. Guida alla società dell'informazione e della credulità*. Roma: Franco Angeli.

Création d'une base de connaissances à partir de messageries spécialisées pour améliorer l'exploration et l'archivage des méls

Touria Ait El Mekki, Bénédicte Grailles, Tsanta Randriatsitohaina

University of Angers – touria.aitelmekki@univ-angers.fr benedicte.grailles@univ-angers.fr tsanta.randriatsitohaina@univ-angers.fr

Abstract
The Pêle-Mél project, supported by the French Ministry of Culture and currently underway, aims to provide archivists with assistance in reading, interpreting, and fine-tuning the context of mailboxes. Our work is based on the observation of archiving practices and on the current state of research in automatic language processing, particularly in computational terminology and information processing. We propose a method of exploration and contextualization of professional mailboxes to facilitate navigation in these systems. It consists, on the one hand, in building a list of terms and named entities to be validated, which we improve by injecting knowledge from specialized thesaurus or message metadata and, on the other hand, in making a classification of messages available via semantic classes grouped around themes chosen by the archivist.

Keywords : archiving, email, machine learning, classification, terminology, named entity

Résumé
Le projet Pêle-Mél, soutenu par le ministère de la Culture français et en cours de réalisation, vise à proposer à l'archiviste une assistance à la lecture, à l'interprétation, à la contextualisation fine des messageries électroniques. Notre travail s'appuie sur l'observation des pratiques d'archivage et sur l'état actuel des recherches en traitement automatique des langues, notamment en terminologie computationnelle et en traitement de l'information. Nous proposons une méthode d'exploration et de contextualisation des messageries professionnelles devant faciliter la navigation dans ces messageries. Il s'agit d'une part de construire une liste des termes et des entités nommées à valider que nous améliorons en injectant des connaissances issues des thesaurus spécialisés ou des métadonnées des messages et, d'autre part, de mettre à disposition une classification des messages via des classes sémantiques regroupées autour de thèmes choisis par l'archiviste.

Mots clés: archivage, courriel, apprentissage, classification, terminologie, entité nommée

1. Introduction

Le projet Pêle-Mél, en cours de réalisation, vise à proposer à l'archiviste une assistance à l'interprétation et la contextualisation fine des messageries électroniques et la rédaction de synthèses à la demande. Notre démarche intègre des méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels, et de l'intervention humaine dans un processus coopératif, et s'appuie sur les résultats de l'élaboration de l'ontologie du domaine à partir du contenu des messageries et de corpus complémentaires. Outre l'identification des termes et la reconnaissance d'entités nommées, nous proposons une projection entre différents plans permettant de faire le lien entre l'ontologie, les entités nommées et les différentes thématiques (Alghamdi et Alfalqi 2015). La méthodologie proposée prend en compte la totalité des informations des messageries (adresses, sujet, contenu, signature, pièces jointes). Les traitements automatiques permettent de construire un ensemble de connaissances à partir des textes et des ressources disponibles, de guider l'intervention humaine en assurant sa cohérence. Nous explorons comment différents types de ressources peuvent être exploités afin de valider des résultats ou pour spécialiser davantage des données préexistantes. Dans ce cadre, nous proposons une approche qui consiste à construire une liste des termes et des entités nommées que nous validons et que nous améliorons en injectant des connaissances issues des thesaurus spécialisés ou des métadonnées des messages. Après une présentation du contexte, de l'état de l'art et des objectifs (section 2), nous reviendrons sur les corpus et la méthodologie (section 3) puis sur la question de l'analyse de textes appliquée aux messages (section 4). Nous souhaitons montrer que l'état de l'art en terminologie computationnelle permet aujourd'hui d'envisager des outils d'analyse fine de textes beaucoup plus puissants.

2. Contexte et objectifs

Les services d'archives sont compétents pour collecter, sélectionner et conserver les documents et données physiques et électroniques. C'est une compétence réglementaire et une obligation légale. Pour répondre à ce besoin, des missions archives ont été placées au sein des ministères. Elles se chargent d'organiser et recueillir les données pour archivage, notamment pour documenter les processus de décision, et de restituer cette information à la demande. Depuis les années 1990, le passage à grande échelle à une production numérique a amené à développer des solutions d'archivage électronique. À partir des années 2000, la question des messageries électroniques a été incluse dans la réflexion sur la pérennisation (Programme Vitam 2013).

2.1. Enjeux

Dans le travail administratif quotidien, ces messageries sont centrales. Support d'informations stratégiques, elles sont les traces uniques de processus décisionnels. Si des outils d'extraction existent, leur volumétrie et leur pérennisation en silos séparés ne suffit pas à satisfaire les besoins d'un bon archivage, à savoir une appréhension par niveaux de granularité et une contextualisation fine. De fait, la recherche dans les corpus se trouve compromise, chaque messagerie étant maintenue dans un entrepôt isolé. C'est l'expérience de recherches difficiles sur des messageries déjà archivées qui a convaincu la mission archives du ministère de la Santé et des Solidarités, forte d'une solide pratique en matière de collecte d'archives électroniques (200 entrées dont 45 % de messageries, avec des boîtes pouvant aller jusqu'à 70 Go) à participer au projet Pêle-mél, initié par les laboratoires Temos et Leria de l'université d'Angers, et soutenu par le ministère de la Culture. Ce partenariat a permis d'accéder à des corpus de messagerie émanant du cabinet, c'est-à-dire de collaborateur.rices personnel.les choisi.es par le.la ministre, ayant pour mission de le.la conseiller et de l'assister dans la réalisation de l'ensemble de ses missions et jouant un rôle administratif et politique à ses côtés.

2.2. État de l'art

Si la question de la pérennisation des messages a été largement traitée, notamment dans le cadre du programme Vitam [http://www.programmevitam.fr], celle de l'accès est beaucoup moins travaillée et uniquement en langue anglaise, via la reconnaissance d'entités nommées : projet ePADD de l'université de Standford [https://library.stanford.edu/projects/epadd] ou RATOM de l'université de Caroline-du-Nord [https://ratom.web.unc.edu/].

Le traitement automatique des courriels est devenu indispensable depuis l'explosion de la quantité de méls échangés entre les différents services et entités. Plusieurs méthodes ont été mises en œuvre pour les explorer. La catégorisation des courriels a été mobilisée dans le but de les organiser, par exemple pour la détection de spam (Tang et al. 2013). Les méthodes utilisées peuvent reposer sur des règles (Xia, 2020), sur l'apprentissage (Nadjate et al., 2020) ou la combinaison des deux. Il est nécessaire d'évaluer l'efficacité des extracteurs automatiques des termes (Nazarenko et al. 2009). L'analyse du contenu des courriels a aussi été utilisée pour catégoriser les contacts en retrouvant les contacts ayant le même centre d'intérêt (Johansen, 2007) ou en identifiant les contacts appartenant à une même communauté (Tyler et al., 2003). Les interactions par méls pouvant être considérés comme un réseau social, des méthodes ont été mises en œuvre pour établir un réseau de contacts en fonction de leurs interactions. Cette analyse se base sur les statistiques des contributions de chaque contact au sein du réseau (Karagiannis, 2009) ou par l'apprentissage

de l'objectif des méls envoyés, pour informer, enquêter et planifier (Lockerd, 2003). La détection d'évènements à partir des courriels a aussi été prospectée permettant l'identification de ceux (recrutement, discours, événement social) mentionnés dans les textes avec les détails comme la date, l'heure, le lieu. Des méthodes de reconnaissance d'entités nommées (Suárez et al. 2020) sont utilisées pour extraire ces informations qui sont ensuite soumis à validation (Nair et al., 2020).

2.3. Objectifs

Au-delà de l'accès à l'information, notre objectif est de proposer à l'archiviste une assistance à la lecture, à l'interprétation et à la contextualisation des messages pris individuellement, comme des boîtes méls dans leur ensemble et une aide à la rédaction de synthèses à la demande. Par exemple, dans un courriel citant le nom d'un.e autre membre du cabinet, nous devons l'identifier sans ambiguïté et faire le lien avec la fonction occupée à la date du courriel. Ainsi, dans un document mentionnant "la ministre des Solidarités et de la Santé" dans un courriel écrit en 2010, nous devons être capable de fournir l'information qu'il s'agit de Roselyne Bachelot-Narquin et permettre la navigation dans les autres courriels envoyés ou reçus et le réseau de correspondant.es. On cherche aussi à identifier des expressions temporelles et des événements. Ce résultat peut être obtenu via l'injection de connaissances complémentaires et externes au corpus de messages, le recours aux méthodes d'apprentissage automatique pour la classification et la catégorisation des contenus textuels, et l'élaboration de l'ontologie du domaine à partir du contenu des messageries (corps du message, fichiers joints, lien) (Maynard et al. 2008) et des listes de vocabulaire contrôlé externes.

L'objectif final est de proposer à l'archiviste une lecture guidée, un instrument d'orientation permettant de progresser dans la restitution du contenu informationnel par sérendipité et des synthèses pour faciliter la gestion des paquets archivés.

3. Corpus et méthodologie

Pour réaliser ce projet, nous disposions initialement de deux messageries de conseillères et de 8 636 messages reçus ou envoyés entre 2007 et 2011. Le contenu des messages en lui-même est court, avec une rédaction correcte, en moyenne 9 phrases par message, mais le contenu reste répétitif. Nous avons donc complété chaque message par les pièces jointes afin d'enrichir le corpus initial et identifié des sources complémentaires : copies d'annuaires papier et organigrammes natifs électroniques ou numérisés, deux thésaurus soit environ 7 000 descripteurs reliés uniquement par des relations hiérarchiques, et 810 discours prononcés par la ministre entre 2010 et 2012. Les annuaires et

organigrammes permettent de repérer les fonctions occupées par une partie des correspondants au sein du ministère et leur évolution au fil des années, ce qui justifie le temps consacré à leur saisie manuelle.
Nous proposons une méthode composée des principaux modules suivants :
- Un module de prétraitement : découpage de courriels, élaboration de réseaux de contacts, extraction des informations comme la fonction, le rattachement des personnes physiques ou morales à partir des adresses méls complétées par les annuaires et les signatures. Une base de données qui contient l'ensemble de ces informations a été construite, ce qui permet de mettre en évidence les liens entre les personnes, les fonctions et les services et facilite aussi différentes visualisations graphiques.
- Un module d'analyse fine de textes : extraction de termes, d'entités nommées, association termes et entités nommées pour appréhender les informations contenues dans chaque courriel.
- Un module de classification thématique afin de parcourir les messages selon les différents thèmes. Il permettra d'avoir une vue plus globale de l'ensemble des messageries.

4. Analyse de textes

À la différence des corpus usuellement utilisés dans les traitements automatiques des textes, les courriels sont plus courts et contiennent plus d'acronymes et d'entités nommées.

4.1. Reconnaissance des sigles

De nombreuses abréviations n'apparaissent pas dans le dictionnaire. Nous avons extrait les abréviations du corpus, grâce à l'étiquetage morpho-syntaxique proposé par TreeTagger, via l'étiquette *abbr* pour abréviation. Après un rapide filtrage, une liste de 254 abréviations dont 195 existaient dans les thesaurus fournis, a été établie. Les spécialistes du domaine ont validé les mots restants. Tous les acronymes n'ont pas été identifiés de cette manière. La liste a été complétée en exploitant les mots étiquetés org via un rapide filtrage. En tout, près de 400 acronymes ont ainsi été détectés.

4.2. Extraction de termes et des entités nommées

Tout mot ou groupe de mots représentant un concept spécifique d'un domaine est un terme. Leur extraction automatique peut être effectuée par des méthodes linguistiques reposant sur les catégories grammaticales, ou par des méthodes statistiques reposant sur les fréquence et distribution des mots dans le texte, ou encore par la combinaison des deux (TermSuite, Cram et Daille, 2016). Les sorties proposent des candidats-termes à soumettre à validation.
Les entités nommées sont généralement assimilées aux noms propres (noms de

personne physique ou morale), noms de lieu ou encore à des valeurs (date, heure...). La reconnaissance d'entités nommées peut être effectuée à partir des bases de connaissances ou par méthode d'apprentissage.
Parmi les outils qui traitent le français et qui répondent au besoin de confidentialité des messageries, nous avons testé plusieurs bibliothèques et choisi la bibliothèque SpaCy (Honnibal et Montani, 2017).
La validation des sorties est exposée à la section 4.4.

4.3. Association des termes et des entités nommées

Les avancées récentes de l'apprentissage automatique ont permis de mettre en œuvre des méthodes permettant la représentation des mots en fonction de leur contexte. On parle de plongement lexical. Celui-ci capture les similarités sémantiques et syntaxiques des mots en fonction du contexte dans lequel ils apparaissent dans le texte. Pour notre projet, nous avons utilisé Fasttext (Bojanowski et al., 2017) afin d'identifier les termes ayant une similarité élevée et retrouver les termes proches des personnes ou des organismes mentionnés dans les textes. Le terme *Solidarités* a été ainsi associé avec *cohésion sociale*, ministre*, Bachelot Narquin, Varenne, direction générale*, autant de désignations cohérentes avec le ministère de Roselyne Bachelot-Narquin. Nous avons pu aussi récupérer des termes associés au nom *Bachelot* comme *ministère*, ministre*, département*, sociale*, *DGCS**, liés à son poste, *jeunesse, politique de la famille*, solidarités, prévention*, sécurité* en lien avec des politiques publiques. Les termes avec * sont présents dans les thesaurus. Après filtrage, les résultats sont prometteurs mais pas suffisants. Sur les 7 000 descripteurs du thésaurus, seuls 60 ont des relations.
L'étape suivante est l'exploration de méthodes et approches terminologique, directes et indirectes, pour structurer la liste des termes et l'outillage à l'aide d'une méthode qui formalise le regroupement conceptuel, le regroupement à base de patrons et les méthodes non terminologiques.

4.4. Validation des termes et des entités nommées

Les entités nommées et termes extraits automatiquement bénéficient d'une phase de validation. Elle s'organise comme suit : par comparaison avec les descripteurs des thesaurus, puis avec la liste des abréviations validées, enfin en confrontant les noms de personne avec les métadonnées des messages et les annuaires. La méthode présentée à la section 4.3 permet une deuxième étape de validation en gardant les termes et les entités nommées associées, suivant la stratégie de Omrane et al., 2011. Le traitement des métadonnées nous permet ainsi de désambiguïser les mentions d'entités nommées et de retrouver que [Luc] *Chatel* a été secrétaire d'État dans le même gouvernement.

4.5. Classification

Nous avons dans un premier temps développé le module de classification en utilisant les méthodes existantes pour regrouper les courriels selon leurs similarités et différences en appliquant un modèle de clustering KMeans, en donnant en entrée au modèle les courriels représentés par tous les mots qu'ils contiennent, puis uniquement des noms, verbes, adjectifs, adverbes, ou par les termes identifiés afin qu'il les distingue par groupes. Les premiers résultats de cette classification n'ont pas été probants. Les centroïdes des clusters sont très proches. La similarité entre le contenu des courriels permet pas au modèle d'identifier des points suffisamment discriminants pour orienter la classification. De plus, nous travaillons sur des textes très spécialisés pour un domaine sans modèle pré-entraîné adapté.

Nous avons amélioré le processus en réalisant une classification guidée par une liste de 50 thèmes structurés en trois niveaux par les spécialistes, donc en regroupant les classes autour de ces thèmes. Après plongements de mots, nous retrouvons le terme *sida* dans la même classe que *ist, vih, maladie, vhc* etc.; le terme *médicament* est associé à *usage, produit, pharmaceutique, generique, prescrit, tamiflu* etc.

Nous avons ensuite représenté chaque thème par sa classe sémantique et calculé la similarité avec chaque courriel pour l'associer aux thèmes les plus proches. La distribution des messages par rapport aux thèmes est cohérente d'après les experts. La confrontation entre celle-ci et la répartition en cours issue d'une annotation manuelle à partir d'un échantillon aléatoire de messages permettra d'évaluer la qualité des résultats.

5. Conclusion

Quelle méthodologie mettre en place pour aider à explorer les courriels et les contextualiser via une lecture guidée ? Un pré-traitement s'impose pour collecter les métadonnées qui servent ensuite à valider des résultats de l'extraction des termes et des entités nommées, en combinaison avec des modèles de plongements lexicaux permettant de les identifier. Pour avoir une vue plus globale des messageries, une classification des messages est indispensable et nécessite des temps de validation par les spécialistes. Nous avons proposé à l'archiviste un nouveau mode d'exploration des messageries conservés à partir des termes, de leurs variants, des classes. Les premiers résultats sont encourageants. La méthode proposée peut être déployée dans d'autres environnements d'archivage.

Bibliographie

Alghamdi R. et Alfalqi K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, vol. 6., pp. 147-153.

Bojanowski P., Grave E., Joulin A. et Mikolov T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, vol. 5, pp.135-146.
Cram D. et Daille B. (2016). Terminology Extraction with Term Variant Detection. In Proceedings of ACL-2016 *System Demonstrations*, pp. 13-18.
Honnibal M. et Montani I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
Johansen L., Rowell M., Butler K. R. et McDaniel P. D. (2007). Email Communities of Interest. In *CEAS The Fourth Conference on Email and Anti-Spam*, 2-3 August 2007, USA.
Karagiannis T. et Vojnovic M. (2009). Behavioral profiles for advanced email features. *In Proceedings of the 18th international conference on World Wide Web*, pp. 711-720.
Lockerd A. et Selker T. (2003). DriftCatcher: The Implicit Social Context of Email. In *INTERACT'03*, pp. 813-816.
Maynard D., Li Y. et Wim P. (2008). NLP Techniques for Term Extraction and Ontology Population. In *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. IOS Press, NLD, pp. 107-127.
Nadjate S., Adi K. et Allili M. (2020). Semantic Representation Based on Deep Learning for Spam Detection. *Foundations and Practice of Security*, pp. 72-81.
Nair A. M., Justus A. A., Ramesh A. et Rajan B. (2020). Event Extraction from Emails. *International Journal of Computer Applications*, vol. 176, n°41, pp. 1-8.
Nazarenko A., Zargayouna H., Hamon O. et Puymbrouck J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *Revue TAL, ATALA*, vol. 50, pp. 257-281.
Omrane N., Nazarenko A. et Szulman S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. *In Proceedings of Conférence internationale de Terminologie et Intelligence Artificielle*, Paris, pp. 80-86.
Programme VITAM (2013). *L'archivage des messageries électroniques. Preuve de concept*, 103 p.
Suárez P., Dupont Y., Muller B., Romary L. et Sagot B. (2020). Establishing a New State-of-the-Art for French Named Entity Recognition. *Language Resources and Evaluation Conference LREC*.
Tang G., Pei J. et Luk W. S. (2014). Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems*, vol. 41, pp. 1-31.
Tyler J.R., Wilkinson D.M. et Huberman B.A. (2003). Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. In Huysman M., Wenger E., Wulf V. (eds) *Communities and Technologies*. Dordrecht, pp. 81-96.
Xia T. (2020). A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems. *IEEE Access*, vol. 8, pp. 82653- 82661.

Order constrained clustering with local stopping rules. Application in textual analysis

Ramón Alvarez-Esteban[1], Mónica Bécue-Bertaut[2]

[1]University of León – ramon.alvarez@unileon.es
[2]Universitat Politècnica de Catalunya – monica.becue@upc.edu

Abstract
The documents of a corpus can be related to an order variable, for example the chronology. This order may explain the variability in the vocabulary. When the vocabulary is highly related to order, correspondence analysis (CA) visualizes the document trajectory on the first factorial axes. Further, hierarchical contiguity constrained clustering methods (HCCC), starting from the document coordinates on these axes, will group the documents only if they are contiguous. However, HCCC may group heterogeneous documents, in terms of their content, because over-emphasis is given to order. To obtain a partition respecting the thematic differences, we propose to include in the algorithm a rule which stops the fusion between two nodes when considered as heterogeneous, meaning that a discontinuity has been detected. HCCC with local stopping rule is applied to Spanish political speeches to illustrate how it works in textual analysis.

Keywords: Hierarchical contiguity constrained clustering, local stopping rule, textual clustering.

1. Introduction

Frequently, the corpus is made up of a series of ordered documents, either according to their production date, the age of the respondent in the case of open-ended questions or another order variable. Hereafter, we use the term chronology as generic. No matter what the chronological corpus is, new words appear, others disappear and the frequencies of the most of them change over time.

CA proves to be a fundamental tool for displaying lexical similarities between speeches and so identifying chronological relationships. When the chronology is highly related to the vocabulary, the first axes reflect the evolution of documents.

However, measuring the importance of the chronology on the corpus structure is not easy when the number of axes is high (high number of documents and words), since the chronology can be distributed among several different factorial axes.

In order to solve this problem, the contribution of hierarchical contiguity constrained clustering method (HCCC) is discussed in Section 2. Section 3 details a stopping rule to avoid grouping together heterogeneous contiguous documents. Section 4 illustrates this procedure in textual analysis by means of an application to investiture Spanish speeches.

2. Cluster analysis and contiguity constrained clustering methods

Hierarchical clustering (HC) documents can use as input data the coordinates of the documents as placed in a Euclidean space by CA. Differences among the clusters can correspond to thematic variations over time. If the chronology effect is strongly dominant, the nodes of the dendrogram built by the HC will group only contiguous elements.

To force the effect of chronology to be taken into account, we can resort to hierarchical contiguity constrained clustering methods (HCCC; Lebart, 1978, Legendre and Legendre, 2012, Borcard et al., 2018, Gançarski et al., 2020), being the contiguity defined by the chronology. HCCC only groups contiguous nodes. Generally, the dendrograms built from both HC and HCCC exhibit differences due to the evolution of the vocabulary over time. In any case, it is very important to understand the information provided by these differences in relation to the results provided by CA.

3. Local stopping rules in contiguity constrained clustering

Local stopping rules (Milligan and Cooper, 1985, Gordon, 1999, Arbelaitz et al. 2013) have been applied in HC with different objectives such as to determine the number of clusters or check their stability.

HCCC can group together chronologically close but heterogeneous nodes. Therefore, it is necessary to determine whether the fusion between two nodes based on their contiguity lead to a heterogenous new node, in the sense that the lexical contents are significantly different. To prevent inversions when applying contiguity-constrained aggregation methods, it is preferable to use the complete-linkage aggregation one (Nielsen, 2016).

To avoid building heterogeneous clusters, we adopt the solution proposed in Legendre et al. (1985) and in Legendre and Legendre (2012) under the name of chronological clustering. Originally, it was designed to detect discontinuities in multi-species time series. In this algorithm, regrouping two contiguous nodes is stopped if the relationship between their contents (computing the distances between all the elements of the two groups) appears to be random depending on a permutation test. The result is a partition of the corpus into homogeneous contiguous documents. A complete hierarchy is not built because of the unauthorised regroupings. However, it is interesting to study the sub-dendrograms (separate parts of the tree) to follow how documents come together.

In some cases, inversions are observed in these dendrograms, even if complete-linkage aggregation method has been used.
Note that the dendrograms obtained with and without using the stopping rule differ and not only because of the discontinuities.

4. Application on a corpus of Spanish investiture speeches

4.1. The corpus

To show the relevance of the local stopping rules on chronological clustering, we use the 16 investiture speeches given by 7 candidates for Prime Minister (Suárez, Calvo-Sotelo, Felipe González, Aznar, Zapatero, Rajoy and Sánchez) from December 1978 to the present day. In their speeches, the candidates present their government programme. Then, the deputies give their confidence or not.
Here, a speech is identified by its rank (from 1 to 16), two letters summarizing the name of the candidate and, finally, the two last figures of the year. So, we have a short number of documents with a similar structure to ease the presentation of the methodology.
The texts of the speeches have been downloaded from the *Diario de Sesiones del Congreso de los Diputados* (Journal of Sessions of the Congress of Deputies). Their writing has been unified (e.g., *IAE* and *Impuesto de Actividades Económicas*). The stopwords are suppressed and only the words with a frequency over 9 and appearing at least in two speeches are retained. The size of the corpus is 162,796 occurrences before applying the selection and 55,850 occurrences after. A total of 1,558 different words are retained.

4.2. Correspondence analysis

After performing CA, 5 axes associated to the above-average eigenvalues (equal to 0.153, 0.094, 0.087, 0.073 and 0.064, respectively) have been retained. All together, they keep 51.4% of the total variance (respectively, 16.7; 10.3; 9.5; 7.9; 7.0). The chronology is indicated here by the rank of the speeches in the series (from 1 to 16), not the year.
Figure 1 represents the speeches on the first two CA planes, (1,2) and (3,4). In this case, the temporal changes unfold along the first three axes. The second axis shows how the first four speeches evolve, while the first axis describes the changes from 4.Gz86 to 12.Sa16, drawing a rough parabola, a usual shape when vocabulary evolves over time. Finally, axis 3 informs about what is characterizing the last four speeches, which is the opposition between 13.Rj16 and 14. Rj16, at the negative part of this axis, and 15.Sa19 and 16.Sa20, on the positive part. It can be assumed that it is an ideological opposition between the candidates of two opposing parties. They are competing for the head of government by means of motions of censure and election repetitions. As our aim is to present the method and not to analyse these speeches in detail, we do not say more here. Let

us just add that both axes 4 and 5 underlines the particularity of the speech 12.Sa16 in relation to those close in time.
This complex configuration will benefit from a synthesis made through clustering starting from the coordinates of the speeches on the first five axes.

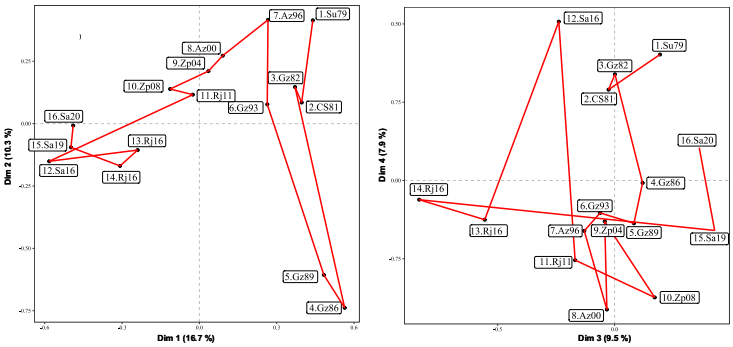


Figure 1. Representation of the documents on the CA planes (1,2) and (3,4).

4.3. Contiguity constrained clustering with local stopping rules

As stopping rule, we adopt here the one proposed in Legendre et al. (1985) and Legendre and Legendre (2012). These works propose to apply a permutation test in order to check whether two nodes, candidates for merger, are homogeneous (merger allowed) or heterogeneous (merger not allowed) at each clustering stage. Depending on the critic p-value (denoted α) that is chosen for this test, different dendrograms are obtained (Figure 2).

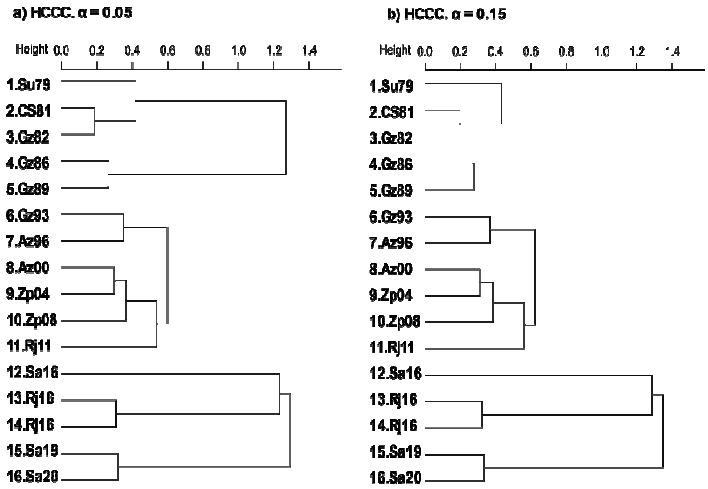


Figure 2. HCCC with stopping local rule with (a) α = 0.05 and (b) α = 0.15.

The number of discontinuities detected in the corpus, and then the number of clusters which constitute time periods, depends on the chosen α value. For α=0.01 (dendrogram not reproduced here) two main clusters would be obtained, from the first speech until 11.Rj11 and from 12.Sa16 to the end, respectively. If we choose α=0.05, three periods are identified (Figure 2.a). The first is composed of the first five speeches from 1.Su79 to 5.Gz89, the second of the speeches from 6.Gz93 until 11.Rj11 and the third of 12.Sa16 and the subsequent ones. When increasing α until 0.15 (Figure 2.b.), the first period is split into two, remaining the others equal. Note that the shape of the dendrograms varies as the α value changes. Legendre et al. (1985) perform several analyses on non-textual data varying the α value between 0.01 and 0.25, allowing more fusions when α is small, leading to larger groups. In general terms, this relationship between the α value and the number of clusters follows this tendency, although exceptions can be found.

After selecting the value of α and thus the partition into clusters, the lexical content of each cluster-period has to be identified looking for the overrepresented and under-represented words in each period. Looking at the most characteristic words of each of the 3 periods, we obtain:

- *Period 1.Sz79 until 5.Gz89*

This first period is characterised by words related to the transition (from dictatorship to democracy), the Socialists' arrival in power, the structure of the State of the Autonomous Regions, the foreign policy, the consolidation of Spain in the European Union as well as the entry into NATO. In particular, the following words are overused in this period: *politics, freedom, problem(s), inflation, unemployment, crisis, balance of payments, infrastructure*. This last one is used in the expressions *"infrastructure development", "infrastructure policy", "infrastructure investments", "infrastructure upgrading"* and so on).

- *Period 6.Gz98 until 11.Rj11*

In this second period, we find many words related to economic growth and development such as: *reform(s), activity, competitiveness, efficient, stability, boost, liberalisation of enterprises* and so on. In particular, the economic crisis gives rise to many words such as: *euro(s), million, financial, debt, deficit, public spending, profit, austerity, economic* and so on.

- *Period 12.Sa16 until 16.Sa20*

This last period is the most complex. Note that this cluster appears in this chronological clustering but also in the case of performing hierarchical clustering without restriction and with contiguity restriction, but without stopping rules. This points that pronounced thematic differences exist between this last period and the former ones. Particular themes are the need to be voted on by a sufficient number of MEPs (*vote(s), agreement, dialogue, pact* and so on), the territorial integrity and the problem posed by Catalonia. New themes are emerging such

as *climate change, corruption crimes, democratic regeneration* as well as equality question with words such as *exclusion, equality, diversity* and *disability*. These themes characterize the entire period.

CA informed us previously that the two speeches by Rajoy of this period, on the one hand, and the three by Sánchez, on the other, are strongly opposed on the third axis (Figure 1). Thus, both present clear lexical differences on which we must also focus on if we want to capture all elements of the corpus structure. Rajoy's speeches insist on words such as *investiture, elections, Partido Popular, Spain, Spaniards* while Sánchez's speeches mention current challenges through the words *digital, digitalisation, ecological* and also more specific problems of left-wing parties such as *workers, poverty, precariousness, vulnerable, inequality, gender* and *women*. Note that this opposition is not visible in clustering results, because the vocabulary that unites the whole period is very important and well differentiated from this of the former periods This fact highlights the complementarity between clustering and CA. The first method captures well the consolidated phenomena while CA is able to detect finer and more subtle aspects.

5. Conclusions

The use of chronological clustering, that is, a particular contiguity constrained hierarchical clustering with stopping rules, has led to detect discontinuities in the corpus and thus determine time periods. Varying the α value leads to different sequencing, some finer than others. Note that an α=0 is equivalent to performing HCCC without stopping rule, considering that all clusters can be considered homogeneous.

In this work, clustering method starts from the coordinates of the documents on the first CA axes. This joint use of CA and clustering allows us to simultaneously interpret CA and clustering results. This is fundamental because the finest aspects of the corpus structure can be hidden in the clustering results.

Software note

Xplortext R package (Bécue-Bertaut et al., 2022), available on R CRAN, has been used for the analysis. Xplortext web offers additional information on this package, as well as examples covered in detail.

References

Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J.M. and Perona I. (2013). An extensive comparative study of cluster validity indices, *Pattern Recognition*, Volume 46, Issue 1, pp. 243-256,

Bécue-Bertaut M., Alvarez-Esteban R, Sánchez-Espigares J.A. and Kostov B. (2022). Xplortext: statistical analysis of textual data. R package version 1.5.

Borcard D., Gillet F. and Legendre P. (2018). Numerical ecology with R, 2nd edition. Use R! series, Springer International Publishing AG. xv + 435 pp.

Gançarski P., Dao TBH D., Crémilleux B., Forestier F. and Lamper T. (2020). Constrained clustering current and new trends. *A guided tour of AI research*. Springer. hal-02548212.
Gordon A. D. (1999). Classification. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.
Lebart L. (1978). Programme d'agrégation avec contraintes [C.A.H. Contiguité], *Les Cahiers de l'Analyse des Données*, 3, 275–287.
Legendre P. and Legendre L. (2012). *Numerical Ecology* (3nd ed.), Elsevier Science BV, Amsterdam.
Legendre P., Dallot S. and Legendre L. (1985). Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton, *American Naturalist*, 125, 257–288.
Milligan G. W. and Cooper M.C. (1985). An examination of procedures for determining the number of clusters in a dataset. *Psychometrika* 50: 159–179.
Nielsen F. (2016). Hierarchical clustering. In: *Introduction to HPC with MPI for data science*. Undergraduate topics in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8

Text Summarization of a scientific document: a comparison of Extractive unsupervised methods

Massimo Aria[1], Corrado Cuccurullo[2], Luca D'Aniello[1], Michelangelo Misuraca[3], Maria Spano[1]

[1]University of Naples Federico II – massimo.aria@unina.it luca.daniello@unina.it maria.spano@unina.it
[2]University of Campania "Luigi Vanvitelli" – corrado.cuccurullo@unicampania.it
[3]University of Calabria – michelangelo.misuraca@unical.it

Abstract
The explosion of scientific literature and the availability of online bibliographic databases makes it extremely difficult for research communities to retrieve rapidly relevant information. Text Summarization aims to condense a given input text into a shorter version generating a reliable summary with the most relevant information about its content. Text Summarization methods can be classified into Extractive and Abstractive. This study computed three extractive unsupervised methods (TextRank, LexRank, and LSA) in order to find keysentences from a scientific document. Our goal was to compare these different approaches to evaluate which method is the most suitable for a scientific full-text.

Keywords: Knowledge synthesis, Text Summarization, Extractive methods, Scientific documents

1. Introduction

The phenomenon of information overload characterized by unstructured textual data is becoming more and more obvious (Sarker et al., 2017). To date, this massive information is mainly provided by the growth of scientific literature, generated every day by research communities spread all over the world. The COVID-19 pandemic marked the biggest explosion of scientific literature ever. Since January 2020, documents on this topic have reached more than 23,000 papers and are doubling every 20 days (Brainard, 2020).

The amount of scientific literature makes it extremely difficult for scholars to select the relevant information they need (Zhang et al, 2022). Due to this, Text Summarization (TS) could represent an innovative solution to this dilemma.

TS is a Natural Language Processing (NLP) technique to find relevant information from one or more documents automatically. It condenses an input text to a shorter version, reduces its size, but at the same time preserves the main

ideas of the input document in less space (Radev et al., 2002) reducing repetition and redundant information to a minimum (Moratanch and Chitrakala, 2017). TS methods help scholars to save time and resources and quickly find the specific key information they are looking for within the documents.

TS methods can be classified as Extractive and Abstractive (Lloret and Palomar, 2012). The Extractive Summarization (ES) carries out a set of the most significant sentences from a document, exactly as they appear. Abstractive Summarization (AS) is similar to the way that a human detects and summarizes the main information of a document. It analyzes the document's text through advanced NLP techniques and generates a summary conveying the most critical information by creating new sentences.

In addition, TS methods can be categorized into supervised and unsupervised. Concerning the availability of prior knowledge, supervised methods (Cheng and Lapata, 2016; Nallapati et al., 2017a) require training data that classify sentences' text as candidates for summary or non-summary. Instead, the unsupervised methods rank sentences with respect to various related features such as term frequency, sentence position, and sentence length (Erkan and Radev, 2004; Filatova and Hatzivassiloglou, 2004). According to these features, a score is assigned to the sentences.

Most automatic text summarization systems are extractive (Ozsoy and Alpaslan, 2011) because building abstract summaries through AS approaches represents a difficult task that involves complex language modeling and requires advanced NLP methods. AS generates an entirely new text from the input document that may not be easily interpreted. Moreover, it is more time-consuming from a computational viewpoint than the ES. Due to these, this study aims to detect and find the most relevant sentences from a scientific full-text document automatically by performing ES unsupervised methods. In order to better capture the meaning component of the full-text, we applied the ES techniques to each document's section separately.

2. Methodology

The extractive process of the keysentences on a scientific full-text was conducted using three types of ES unsupervised methods in order to investigate, compare, and evaluate their findings. The three ES techniques are TextRank, LexRank, and Latent Semantic Analysis (LSA). TextRank (Mihalcea and Tarau, 2004), based on Google's PageRank algorithm (Page et al., 1999), is a word graph-based ranking model. The algorithm models any document as a graph using sentences as nodes. The result of this process is a dense graph representing the document. Then, the PageRank algorithm is performed to identify the most relevant sentences.

To find relevant sentences, the TextRank algorithm needs 2 inputs: a data frame

with sentences and a data frame containing tokens included in each sentence. A similarity measure between every pair of sentences is needed to build edges. This function is used to weight the graph edges, the higher the similarity between sentences the thicker the edge between them will be in the graph. We used the Jaccard index (Jaccard, 1912) to measure the similarity between sentences and, consequently, to count the percent of common tokens.

LexRank (Erkan and Radev, 2004) is another graph-based ranking model. The algorithm is programmed to summarize a set of documents similar in terms of topics by proposing which sentences share the most information. As TextRank, a graph is built representing sentences as nodes and similarity between them as edges. *The modified idf cosine* similarity (Erkan and Radev, 2004) is used to measure the similarity between pairwise sentences. A minimum similarity threshold for a sentence pair needs to be fixed to avoid unimportant links in the graph. Then the sentences are ranked according to their similarities by using their degree of centrality or the PageRank algorithm. The advantage of this algorithm is that clusters share similar content and reflect the different topics of the collection. From each cluster, a centroid sentence is obtained and work as synthesis of the content for all other sentences in the cluster.
In our study, LexRank was run on a single scientific document where the cluster of documents is represented by the sentences of each document section. To compare findings with TextRank, we used the PageRank algorithm for ranking sentences.
LSA (Gong and Liu, 2001) works differently with respect to the ES methods described above. It is used to find hidden semantic structures of words and sentences. LSA formalizes semantic word representation within a vector space, typically named semantic space, whose dimensions have been reduced using an algebraic method called Singular Value Decomposition (SVD – Deerwester *et al*., 1990). SVD allows finding the interrelations among words and sentences identifying the patterns of relationships between terms and concepts. It has the capability of reducing noise reduction, which helps to improve accuracy (Ozsoy and Alpaslan, 2011). LSA is based on the assumption that words that occur in the same contexts tend to have similar meanings. After representing the input text as a matrix and calculating SVD values, a matrix *Concepts x Sentences* is obtained. Each cell of this matrix represents the relationship between sentences and concepts. A higher value indicates that the sentence is more related to that concept. Since concepts are ordered with respect to the explained variability, the algorithm chooses one sentence from the most important concept, then a second sentence from the second most important concept, and so forth. In this study, we extracted three sentences, each of them representative of the first three concepts. All the summarization analyses were run in the R environment (v. 4.2.0 – R Core

Team, 2016), using texkrank (Wijffels, 2020), lexRankr (Spannbauer and White, 2019), LSAfun (Guenther et al., 2015) packages.

3. Data collection and Main findings

This study aims to compare the three ES methods in each section of a scientific full-text document. In particular, we consedered a paper published in the Social Indicator Research journal titled "Mapping the Evolution of Social Research and Data Science on 30 Years of *Social Indicators Research*" (Aria *et al.*, 2020). In this paper, the whole collection of publications indexed on SIR from 1989 to 2018 is reviewed, providing a complete overview of the main factors that characterized the life of the journal in the last 30 years. The theoretical framework of bibliometric studies was used to analyze this enormous collection of documents. We choose to analyze the full-text of this paper because it was published by three of the authors who contributed to this study. In this way, the authors were able to check the reliability of findings obtained from the ES methods. Each paper's section was analyzed and pre-treated separately. The text was split into sentences by considering punctuation marks such as dots and semi-colons. Then, a tokenization process was performed, obtaining a set of distinct strings, namely tokens (i.e., the words used in the text) separated by blanks. However, since the expression "*word*" is too generic and does not encompass other valuable combinations of words (e.g., collocations and multiword expressions) *term* is conventionally used instead of *word* (Misuraca and Spano, 2018). Data were pre-treated to reduce language variability and avoid possible sources of noise, improving in this way the effectiveness of the next analytical steps (Song et al., 2005; Uysal and Gunal, 2014). Because of case sensitivity, all the characters of terms were translated in lower case. Moreover, other normalizing operations were applied such as deleting numbers, references cited in brackets, and dots used for an abbreviated author's name. Then, an ID was assigned to each sentence.

The main findings of the analysis are summarized in Table 1. On the left side (A), we reported the main information about the structure of the paper by focusing on each section. The paper counts 7 sections. The length of each section can be measured by the number of sentences and, consequently, by the number of terms counted after removing non-informative terms (e.g., *stop-words*). These measures allow identifying which parts of the paper the authors were mainly focused on. Sections 3, 4, and 6 are the longest. They represent the core themes of the paper, talking about: the data retrieval process and the inclusion and exclusion criteria used for the selection of the documents (Section 3); the methods used to measure the scientific productivity of the authors, and their collaborations (Section 4); the use of science mapping techniques to analyze the conceptual structure of the collection from a longitudinal viewpoint (Section 6).

On the right side (B), the top 3 keysentences extracted with the ES methods are reported. The common keysentences are in bold. TextRank and LexRank have more common keyphrases (in sections 1,4,5, and 7) than LSA. The keysentences are the same because both the algorithms work in a similar way. LSA has just four keysentences in common with the ones of the other two ES methods. The approach of Gong and Liu has some disadvantages that are defined by Steinberger and Jezek (2004). The first disadvantage is that if the given predefined number is large, sentences from less significant concepts are chosen. Moreover, the second disadvantage is related to choosing only one sentence from each concept. As we said above, concepts are ordered with respect to their importance and some of them can contain sentences that are highly related to the concept but do not have the highest value. The last disadvantage is that all chosen concepts are assumed to be at the same importance level, which may not be true.

Table 1. Information about Sections' structure (A) and Top 3 keysentences of ES methods (B)

Sections' info (A)				**Top 3 keysentences (B)**		
ID	***Title***	***# Sentences***	***# Terms (Mean ± SD)***	***TextRank***	***LexRank***	***LSA***
1	Introduction	38	466 (12.3 ± 5.58)	**14**;**1**;9	**1**;**26**;**14**	**26**;15;24
2	Bibliometric Methodology and Related Work	27	336 (12.4 ± 4.1)	18;9;15	17;**1**;19	3;12;**1**
3	Social Research and Data Science on 30 Years of SIR (1989–2018)	66	701 (10.6 ± 4.69)	19;29;50	**44**;15;47	**44**;48;36
4	Productivity and Collaboration Networks of Countries and Institutions	74	747 (10.1 ± 3.87)	**6**;36;34	**6**;29;38	39;50;27
5	Highly Contributive Authors, Papers and Citations	53	537 (10.1 ± 3.83)	**1**;**10**;12	32;**1**;**10**	4;24;3
6	Thematic Analysis and Evolution	71	780 (11 ± 4.63)	40;35;36	34;57;54	10;12;5
7	Conclusions	37	382 (10.3 ± 5.59)	**2**;**27**;30	**2**;17;6	**27**;14;3

4. Final remarks and conclusions

Finding out the relevant information from huge scientific documents is becoming more and more necessary. Text summarization, with its ES unsupervised methods, could be the key solution to consult retrieved documents and decide on their relevance rapidly.

TextRank and LexRank seem to run better than LSA when the input text is a scientific full-text. However, some remarks need to be highlighted. An algorithm for data retrieval, data import, and data pre-treatment needs to be developed. It can be programmed with advanced regular expressions to allow recovering the full-text by considering different journal templates. Then, other methods for computing the input matrix or similarity measures for sentences can be developed to improve ES findings. Indeed, in future work, the *modified tf-idf* proposed by Ozsoy and Alpasla (2011) can be performed in order to eliminate noise from the input matrix. Another improvement can be performing ES methods to find the most informative citation in a scientific full-text (the In-Citation Summarization). Finally, evaluates the ES by considering multiple scientific full-texts as input.

References

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(3), 803-831.

Brainard J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?, 2020-05. *Science*. American Association for the Advancement of Science (AAAS). doi: 10.1126/science.abc7839.

Cheng J. and Lapata M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (Volume 1: Long papers) (pp. 484–494). Berlin, Germany: Association for Computational Linguistics.

Erkan G. and Radev D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22, 457-479.

Filatova E. and Hatzivassiloglou V. (2004). Event-based extractive summarization. In *Text summarization branches out* (pp. 104–111). Barcelona, Spain: Association for Computational Linguistics.

Gong Y. and Liu X. (2001) Generic text summarization using relevance measure and latent semantic analysis. In: *Proceedings of SIGIR '01 2001.*

Guenther F., Dudschig C. and Kaup, B. (2015). LSAfun: An R package for computations based on Latent Semantic Analysis. Behavior Research Methods, 47(4), 930-944. doi:10.3758/s13428-014-0529-0.

Lloret E. and Palomar M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1), 1-41.

Mihalcea R. and Tarau P. (2004) TextRank: Bringing Order into Texts.In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp. 404–411.

Misuraca M. and Spano M. (2018). Unsupervised analytic strategies to explore large

document collections. In *International Conference on the Statistical Analysis of Textual Data* (pp. 17-28). Springer, Cham.
Moratanch N. and Chitrakala S. (2017). A Survey on Extractive Text Summarization. Paper presented at the 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai.
Nallapati R., Zhai F. and Zhou B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 3075–3081). AAAI Press.
Ozsoy M. G., Alpaslan F. N. and Cicekli I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4), 405-417.
Page L. Brin S. Motwani R. and Winograd T. (1999) The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, pp. 1–17.
Radev D. R., Hovy E. and McKeown K. (2002). Introduction to the special issue on summarization. Computational Linguistics, 28(4), 399-408. doi: 10.1162/089120102762671927.
Sarker A., Molla D. and Paris C. (2017). Automated text summarisation and evidence-based medicine: A survey of two domains. arXiv preprint arXiv:1706.08162.
Song F, Liu S and Yang J (2005) A comparative study on text representation schemes in text catego- rization. Pattern Analysis and Applications 8(1–2):199–209.
Spannbauer A. and White B., (2019). lexRankr: Extractive Summarization of Text with the LexRank Algorithm. R package version 0.5.2. https://CRAN.R-project.org/package=lexRankr.
Uysal A. and Gunal S. (2014) The impact of preprocessing on text classification. Information Processing and Management 50(1):104–112.
Wijffels J. (2020). textrank: Summarize Text by Ranking Sentences and Finding Keywords. R package version 0.3.1. https://CRAN.R-project.org/package=textrank.
Zhang C., Zhao L., Zhao M. and Zhang Y. (2022). Enhancing keyphrase extraction from academic articles with their reference information. *Scientometrics*, 127(2), 703-731.
Jaccard P. (1912) The distribution of the flora in the alpine zone, 1, *New phytologist*, 11(2), 37-50.
Deerwester S., Dumais S., Landauer T., Furnas G. and Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391-407.

Towards Automatic Screening for Fibromyalgia in Italian Social Media Users

Valerio Basile

University of Turin – valerio.basile@unito.it

Abstract
Fibromyalgia (FM) is a syndrome characterized by a number of symptoms including chronic pain, tiredness, and cognitive dysfunctions. Medical studies estimate a widespread incidence of FM, severely skewed towards women. However, while the European Parliament recognizes FM as a condition negatively impacting the lives of millions, and despite estimates of about 2 million people suffering from FM in Italy, the condition is treated unevenly across this country. One of the main obstacles toward full healthcare for FM patients in Italy is the difficult and often excessively long diagnostic path.
The goal of this study is to leverage the vast amount of natural language data available in social media, in order to model the language of FM and build an automatic system that distinguishes users suffering from FM from healthy users based on their social media post history. To this aim, we collected about 250K messages from Twitter, in Italian, from 145 users who declare to suffer from FM, and an equivalent amount of messages from random users as a control group. We build supervised classifiers with traditional machine learning techniques, namely Support Vector Machine and Random Forest, obtaining a 72% accuracy in a cross-validation experiment aimed at predicting the user class as FM or not-FM. The classifiers employ explicit features such as ngrams and lemma counts from the Italian translation of Linguistic Inquiry Word Count (LIWC), which provide an interpretable insight into the language of people with FM. We further implemented a state-of-the-art classifier based on AlBERTo, the Transformer model based on BERT pre-trained on a large collection of Italian tweets, bringing the classification accuracy up to over 78%. The high precision (0.82) on the positive class (FM) represents a promising result towards automatic, non-invasive screening of Fibromyalgia on Italian social media users.

Keywords: Fibromyalgia, Twitter, Transformer Network, Supervised Machine Learning

1. Introduction

Fibromyalgia (FM) is a "condition that causes pain all over the body, sleep problems, fatigue, and often emotional and mental distress" per the definition by the United States Center for Disease and Control Prevention (CDC), which adds "people with fibromyalgia may be more sensitive to pain than people

without fibromyalgia". While several treatments are found to work on some of the symptoms (Clauw, 2014), the root cause of FM is still unknown, and the diagnosis is difficult (Goldenberg, 2009). FM is not a rare condition, with a conservative estimate of 2% to 4% of people suffering from FM worldwide, women being more interested than man in a 2:1 ratio (Fitzcharles et al., 2021).
In Italy, FM's estimated prevalence is in 2.22% of the population (Salaffi et al., 2005). Despite the institution of a centralized registry (IFR, Italian Fibromyalgia Registry) and the recent allocation of budget to study FM by the Italian Government in the beginning of 2022, the management and diagnosis of FM patients in Italy is heterogeneous across the 20 Italian regions. FM has a negative impact on the lives of more than two million Italians, many of which are aware of the symptoms but never received a coherent diagnosis.
The goal of this paper is to study the language of people suffering from FM, as used in social media, and in particular the Italian Twittersphere. We collected a large number of microblog posts from self-declared FM patients and created classifiers based on Natural Language Processing techniques, with a two-fold aim: i) to understand the common patterns of topics which are prevalent in the day to day discourse by FM patients; ii) to test the feasibility of an automatic discovery tool to support large-scale screening for FM on Italian social media users.

2. Data

In order to analyze the language of social media users in Italy, we collected a corpus of posts from Twitter, the popular microblogging platform. We accessed the Twitter API on December 5th, 2021, searching for the term "fibromialgia" (Italian spelling of "fibromyalgia") and filtering for the Italian language. We furthermore excluded retweets, replies, and quotes from the query results, as well as tweets containing multimedia elements and links to external pages. The exact query, in the Twitter API syntax, is the following:
(fibromialgia) lang:it -is:retweet -is:reply -is:quote -has:links -has:media -has:images -has:videos -is:nullcast
The *-is:nullcast* option excludes "Promoted-only" tweets, i.e., tweets that, once published, do not appear in the user's public timeline.
The query retrieved 1,596 tweets, that underwent a round of manual annotation, with the goal of identifying users that suffer from FM. A relatively large number of tweets were found to be not written in Italian. More precisely, these tweets were Spanish or Portuguese, due to the identical spelling of "fibromialgia" in these languages and Italian, and the less than optimal accuracy of the Twitter API in terms of language identification. Of the remaining 1,348 tweet, we manually identified 339 tweets from 148 unique authors self-declaring to be suffering from FM. The criteria to decide whether a user is a FM patient or not were strict: the user has to declare explicitly to have received a FM diagnosis or

otherwise be a FM patient, as in this example:
[IT] Salve ho bisogno del vostro aiuto soffro di una patologia rara chiamata fibromialgia aiutatemi per favore
[EN] Hi I need your help I suffer from a rare condition called fibromyalgia help me please
or has to imply it clearly, e.g.:
[IT] se potessi dare un nome al mio dolore oggi lo chiamerei inferno, la fibromialgia a volte mi prende corpo e spirito e mi porta negli abissi.
[EN] if I could name my pain today I would call it hell, sometimes fibromyalgia takes my body and soul and takes me to the abyss.
Hashtags are also considered a signal, as in this example:
[IT] Maledetti, maledetti dolori.... #Fibromialgia
[EN] Damned, damned pain…. #Fibromyalgia
We accessed the Twitter API a second time, and downloaded all the tweets from the 145 users (3 of the authors were not accessible or they had no tweets in their timeline), excluding retweets but including replies to other messages, retrieving 250,922 tweets in total.
In order to build a balanced corpus, we extended this collection with a similar number of tweets from users allegedly not suffering from FM. To this aim, we first sampled 300 tweets from TWITA (Basile et al. 2018), a large-scale collection of tweets in the Italian language available for research purposes. The tweets were selected randomly from the data of December 4th, 2021. We manually annotated this set of tweets to exclude tweets from non-individuals (e.g., organizations), and tweets in languages other than Italian, and finally randomly selected 145 users among the authors of the tweets. From these users, we retrieved 284,313 tweets to balance the original set of tweets from FM patients.

3. Method

We performed a series of classification experiments in order to validate the dataset introduced in the previous section, and test the feasibility of a predictive model for distinguishing FS patients based on their Twitter posts.
We preprocessed the posts, and created three sets of features to represent the users based on their language from the dataset. The preprocessing is carried out mainly with UDpipe 2.5 (Straka et al., 2016), a Natural Language Processing pipeline based on the Universal Dependency theory of syntax (De Marneffe et al., 2021). We equipped UDpipe with the PoSTWITA pre-trained model, based on the only available gold standard treebank of Italian tweets (Sanguinetti et al., 2018). URLs and mentions are replaced with special tokens in the tweets with the *tweet-preprocessor* Python library, and then tokenized with UDpipe, preserving URL and mention tokens.

3.1. Features

The first set of features are TF-IDF-weighted counts of token ngrams of length ranging from 1 to 4. These feature vectors are extracted with the *TfidfVectorizer* method of the *Scikit-learn* Python library, setting the *max_count* parameter to 5,000, i.e., selecting the top 5,000 ngrams by frequency count. Note that the ngrams are counted considering all the tweets from each individual users, therefore the process produces one 5,000-dimensional feature vector per user. The intended goal of this vectorization is to capture the fine-grained lexical knowledge expressed in the tweets by modeling the relevance of individual words and short phrases.

The second set of features is based on the Italian translation of the Linguistic Inquiry Word Count dictionary (Agosti et al., 2007). In this version of LIWC, more than 100,000 Italian words are associated to 85 categories rooted in psycholinguistics. For each post of each user, we counted the words occurring in any of the 85 categories of the dictionary, thus producing a 85-dimensional feature vector for each user, where each dimension corresponds to a LIWC category. The intended goal of this vectorization is to capture higher-level concepts expressed in the text of the tweets, especially those linked to psychological traits of the authors.

The third set of features is the encoding produced by a state-of-the-art neural language model, namely AlBERTo (Polignano et al., 2019), a model based on BERT (Devlin et al., 2019) pre-trained on a large corpus of Italian tweets from TWITA (Basile et al., 2018). BERT is a complex neural network architecture that incorporates a series of *transformer* modules. The network is trained on a large quantity of raw text (Italian tweets in the case of AlBERTo) to solve the unsupervised of *masked language modeling*, i.e., predicting a random subset of words in a sentence, removed ("masked") for the purpose of training, from the remaining context. The resulting language model encodes a great deal of knowledge about the language variety of the training corpus, and it can be *fine-tuned* to solve specific, supervised tasks such as text classification. Without fine-tuning, the neural language model can be directly used to encode unseen instances of natural language. In this work, for each post, AlBERTo generated an embedding in the form of a 768-dimensional dense vector. In order to create user-level embeddings, we averaged all the embedding of their posts.

3.2. Classification experiments

The experiments are ten-fold cross-validations, with a Random Forest classifier (Breiman, 2001), where the unit to predict is a Twitter user from our dataset and the label is binary, i.e., FM patient or not. Table 1 shows the results of the classification with different combinations of features.

Table 1. Results of the 10-fold cross-validation, reported in terms of precition (P.), recall (R.), and F1-score (F1), on the negative class (0), the positive class (1), and their macro-average (m), and accuracy (Acc.). The rows corresponds to the combinations of feature sets (N for ngrams, L for LIWC, BERT for AlBERTo embeddings).

Features	P. (0)	R. (0)	F1 (0)	P. (1)	R. (1)	F1 (1)	P. (m)	R. (m)	F1 (m)	Acc.
N	0.742	0.746	0.737	0.761	0.738	0.743	0.751	0.742	0.740	0.742
L	0.746	0.704	0.715	0.727	0.745	0.728	0.737	0.725	0.722	0.725
N+L	0.754	0.746	0.742	0.763	0.753	0.751	0.758	0.749	0.747	0.749
BERT	0.767	0.820	0.787	0.820	0.746	0.775	0.794	0.783	0.781	0.784

As expected, the combination of ngrams and LIWC counts performs better than either set of features individually. The neural embeddings, even if aggregated in a straightforward fashion by simple average, perform significantly better than both other sets of features. Ngram counts and LIWC, however, are interpretable features, and the Random Forest classifier outputs a value of feature importance, which allows us to investigate more in depth the language of FM patients.
Indeed, the top ngrams ranked by importance are *dolore*, *dolori* (singular and plural form of "pain"), *non lo* (an incomplete negated syntactic structure similar to "I don't… it"), and *il dolore* ("the pain"). It is evident how much chronic pain is continuously mentioned in the social media posts of FS patients, along with an indication of things that they do not want or cannot do. The top features ranked by importance in the classification performed with LIWC featured only are, in decreasing order of importance, Body, Human, Home, Consent, Self, and Physical, suggesting that the language of FS patients is somewhat focused on themselves and their status.

4. Conclusions

In this short paper, we investigated the language of people suffering from Fibromyalgia (FM) through their posts on Italian social media. A predictive analysis on a carefully handcrafted dataset of Twitter users uncovered insights about the condition of self-declared FM patients, such as chronic pain, and showed that it is possible to predict the presence of FM from the natural language alone to a promising extent.
Employing a state-of-the-art transformer-based classifier makes it more challenging to investigate which language features correlate with the FM label. However, the absolute classification performance in this case is significantly higher. In particular, the high precision on the positive class, which could be further tuned, e.g., modifying the threshold on the sigmoid activation, paves the way for a fully automated NLP engine that could run as a bot on Italian Twitter for the purpose of large-scale screening.

The language technology presented in this paper could be further improved by training the model on data from authorized medical sources, with the ultimate goal of improving the well-being of millions of Italian speaking citizens suffering from Fibromyalgia.

References

Agosti A., Rellini A. (2007). The Italian Liwc Dictionary. *LIWC.net*; Austin, TX, USA: 2007. Technical Report.

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.

Breiman L. (2001). Random Forests. *Machine Learning*, 45, 5-32. DOI: 10.1023/A:1010933404324 .

Choy E. et al. (2010). A patient survey of the impact of fibromyalgia and the journey to diagnosis. *BMC health services research*. DOI: 10.102.10.1186/1472-6963-10-102.

Clauw J. (2014). Fibromyalgia: A Clinical Review. JAMA, 311(15), 1547–1555. https://doi.org/10.1001/jama.2014.3266

De Marneffe C., Manning C., Nivre J., Zeman D. (2021): Universal Dependencies. *Computational Linguistics*, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308

Devlin J., Chang M., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Fitzcharles et al (2021). Nociplastic pain: towards an understanding of prevalent pain conditions. *Lancet* 2021; 397: 2098–110

Goldenberg D. L. (2009). Diagnosis and differential diagnosis of fibromyalgia. *The American journal of medicine*, 122(12 Suppl), S14–S21.
https://doi.org/10.1016/j.amjmed.2009.09.007

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. *Proceedings of the Sixth Italian Conference on Computational Linguistics* (CLiC-it 2019). CEUR

Salaffi, F., De Angelis, R., Grassi, W. (2005). Prevalence of musculoskeletal conditions in an Italian population sample: results of a regional community-based study. I. The MAPPING study. *Clinical and experimental rheumatology*, 23(6), 819–828.

Sanguinetti M., Bosco C., Lavelli A., Mazzei A., Antonelli O., Tamburini F. (2018). PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018).

Straka M., Hajič J., Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16): 4290–4297.

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

Federica Beghini

Università degli Studi di Padova,
Université Côte d'Azur (Laboratoire BCL) –
federica.beghini@phd.unipd.it

Abstract
This article presents a study of characteristic morphosyntactic elements of Milan Kundera's novel production, which are detected by linguistic, statistical and machine learning approaches. The specificity of this contribution is to propose, in addition to the traditional statistical methods, a deep learning training on a database comprising Kundera's novels and a representative sample of French contemporary novels, using grammatical categories as sole representation of the texts. This study led to the identification of a morphosyntactic pattern, which was then examined to reveal the aesthetic intention behind its use.

Keywords: French linguistics, stylometry, French literature, deep learning.

Résumé
Cet article présente une étude des éléments morphosyntaxiques caractéristiques de la production romanesque de Milan Kundera, qui sont détectés par des approches linguistiques, statistiques et d'apprentissage profond. La spécificité de cette contribution est de proposer, outre les traditionnelles méthodes d'exploration statistique, un entraînement d'apprentissage profond sur une base comprenant les romans de Kundera et un corpus représentatif des romans contemporains, en utilisant les catégories grammaticales des mots comme seule représentation des textes. Cette étude a permis d'identifier un patron morphosyntaxique, qui a été ensuite examiné pour révéler l'intention esthétique qui en sous-tend l'emploi.
Mots clés : linguistique française, stylométrie, littérature française, deep learning.

1. Introduction

Ces dernières années, le domaine des études textométriques a connu un intérêt croissant pour les approches qui, outre l'étude traditionnelle des unités simples d'un texte (forme lexicale, lemme, catégorie morphosyntaxique, ponctuème), visent également à examiner les relations syntagmatiques que celles-ci forment entre elles. Ainsi, dans les deux dernières décennies, la textométrie a abordé la

question de l'étude des unités complexes, ou, selon les termes de Legallois (2018), « des séquences d'unités », « des schémas de phrase », des « patrons » ou des « motifs ». À cette fin, différentes démarches ont été proposées et explorées, relevant de l'ADT (Legallois, 2018 ; Longrée *et al*. 2008, 2013) et du *deep learning* (Magri 2020 ; Mayaffre et Vanni, 2021). Cette étude s'inscrit dans cette lignée de recherche : son objectif correspond à l'identification des patrons prototypiques de la prose de Kundera, plus précisément des schémas de phrase morphosyntaxiques les plus significatifs de son écriture, par rapport à celle d'un échantillon représentatif de la littérature française contemporaine à notre auteur. Dans ce but, une approche innovante des méthodes du *deep learning* pour l'étude des unités complexes d'une œuvre littéraire a été élaborée, employée conjointement aux outils d'exploration statistique.

2. Une méthodologie intégrée

La notion de « motif », définie par Longrée *et al*. (2008 ; 2013), Legallois (2018), Mayaffre et Vanni (2021), correspond à l'association récurrente de certains éléments textuels, continus ou discontinus, pouvant appartenir à différents niveaux linguistiques (lexique, catégories grammaticales, syntaxe) qui viennent former des configurations ou des patrons lexico-grammaticaux. Dans cette étude, nous ne concentrerons notre analyse que sur le plan morphosyntaxique, à savoir sur les unités complexes les plus saillantes des catégories grammaticales.

Cette recherche porte sur les neufs romans de l'*Œuvre I et II* de Kundera (2017, Gallimard), pour un total de 823.261 occurrences. L'analyse de ce corpus se sert d'un étalon de référence qui a été compilé pour qu'il puisse représenter de façon homogène la langue littéraire française du genre romanesque de la période contemporaine à l'activité littéraire de cet auteur (1960-2019). Ce modèle de référence contient 35 romanciers pour un total de 9.418.136 occurrences.

L'étude des configurations morphosyntaxiques a été menée à l'aide d'une approche *corpus-based* : l'exploration quantitative du corpus s'est basée sur des hypothèses dérivant d'une étude qualitative de l'œuvre de Kundera. Plus précisément, par le biais de l'exploration statistique et de l'apprentissage profond, nous avons mis à l'épreuve nos considérations qualitatives concernant la récurrence d'un certain type de groupe nominal, à savoir [Dét. + N1 + Prép. + Dét. + N2].

Ces analyses quantitatives ont été menées à l'aide du logiciel Hyperbase (version web et version standard) et elles s'appuient à la fois sur l'exploration statistique et sur l'apprentissage profond pour une classification de textes (Savoy, 2015) – plus précisément, une tâche d'attribution – qui utilise des réseaux neuronaux convolutifs (*convolutional neural network*) (Kalchbrenner et al., 2014). En effet, depuis les travaux de Vanni et al. (2018), ces modèles ne se limitent plus à la tâche de classification, mais autorisent également un mécanisme inverse, celui

de la déconvolution, en mesure d'extraire les éléments textuels à la base de cette classification, à savoir les *marqueurs linguistiques* sur lesquels repose son choix. Ainsi, en ce qui concerne le *deep learning*, nous avons effectué autant d'entraînements que le nombre de romans de Kundera à classifier, à savoir neuf, seulement sur les catégories grammaticales. Pour chacune de ces neuf analyses, l'algorithme s'est entrainé sur une base contenant à la fois le corpus de référence et tous les romans de Kundera sauf celui à classifier. Par exemple, si le roman à classifier était *L'Identité*, il était enlevé de la base et, après l'entraînement, il était soumis à l'algorithme pour la tâche de classification ; la même procédure a été employée pour tous les romans de Kundera. Les textes de notre auteur ont tous été reconnus comme étant de lui, avec des taux de précision variant entre 40% et 70%, avec l'exception du premier roman, qui est attribué à Kundera, mais avec un taux de précision plus bas, de 24%. Enfin, tous les résultats de ces neuf tâches de classification ont été examinés et, dans la prochaine section, nous présenterons des marqueurs linguistiques qui reviennent et qui peuvent donc être considérés comme des éléments linguistiques distinctifs de la prose de Kundera par rapport à celle de notre échantillon représentatif de la prose romanesque contemporaine à l'auteur.

3. ADT et *deep learning*. Les configurations morphosyntaxiques

Tout d'abord, nous avons examiné la structure [Dét. + N1 + Prép. + Dét. + N2] à l'aide des méthodes statistiques. Un calcul des spécificités (fig. 1) a confirmé sa significativité dans les romans de Kundera et une surutilisation dans son œuvre par rapport à ce qu'on observe dans la langue littéraire contemporaine, avec un écart réduit de +35.4.

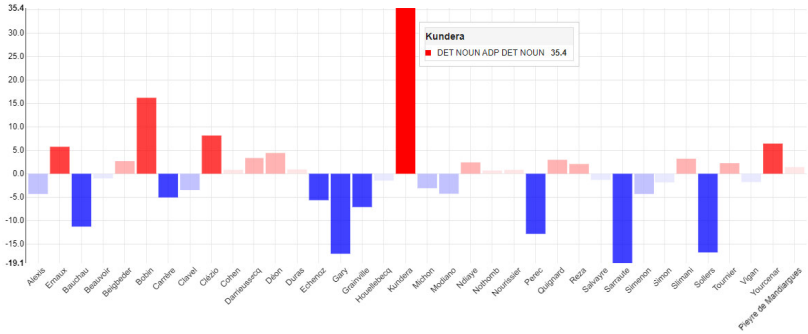


Figure 1. Le schéma [Dét. + N1 + Prép. + Dét. + N2] (Hyperbase – 2022)

Ce « schéma de phrase » a également été relevé dans les classifications du *deep learning* : dans les zones d'activation les plus significatives – c'est-à-dire qui présentent un TDS élevé (Vanni *et al.*, 2018), – le processus de déconvolution a détecté des suites de codes qui le rappellent.

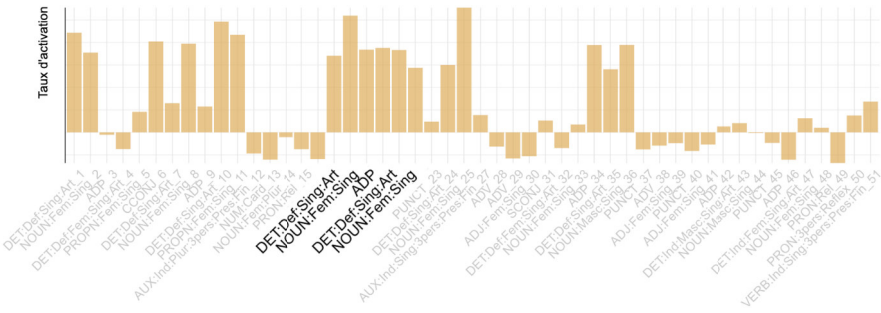


Figure 2. Le deep learning. Le marqueur linguistique [Dét + N + Prép + Dét + N] (Hyperdeep 2022)[1]

Étant donné l'impossibilité d'inclure toutes les sorties des neuf classifications, nous rapportons ici sous forme discursive les marqueurs linguistiques qui, en plus d'avoir un TDS élevé, sont communs aux neuf classifications. La même séquence de composants de la structure que nous venons de relever se succède (fig. 2), bien que souvent seules des parties du patron apparaissent, comme [N + Prép + Dét + N] et [Prép + Dét + N]. Les résultats qui ne se réfèrent pas spécifiquement à cette suite de catégories grammaticales, tels que [N...N] ou [Dét + N... Dét + N], détectent néanmoins la saillance de la succession des substantifs. Enfin, d'autres zones d'activation mettent en évidence des variantes concernant la présence des adjectifs qui précèdent ou suivent les noms ; le cas d'adjectifs antéposés est le plus significatif, comme dans le marqueur [Prép + Dét + Adj + N].

Ainsi, la suite de codes relevée par une étude qualitative du texte et vérifiée par la statistique a également été détectée par l'apprentissage profond. En plus de démontrer sa significativité, le *deep learning* nous a également permis d'en identifier des variantes [Prép + D + *Adj* + N] et de relever la saillance de la catégorie grammaticale du substantif.

3.1. Le retour au texte. À la recherche des motifs morphosyntaxiques

Les sorties du *deep learning* ont orienté notre étude du concordancier, dans laquelle nous avons donc décidé d'inclure des variantes contenant l'adjectif antéposé et postposé. Ce retour au texte par le biais de l'analyse des cooccurrences sert à l'identification des constantes de ce schéma grammatical,

[1] « Pour avoir la certitude que l'amitié érotique ne cède jamais à **l'agressivité de l'amour**, il ne voyait chacune de ses maîtresses permanentes qu'à de très longs intervalles. » (Kundera, 2017, p. 1149).

afin de dégager l'intention esthétique qui en sous-tend l'emploi. Pour des raisons d'espace, nous ne rapportons qu'une sélection des exemples les plus significatifs.

1	Curieuse alliance :	**la froide impersonnalité de la technique**	et les flammes de l'extase.
2	Imprimer une forme à une durée, c'est	**l'exigence de la beauté,**	mais aussi celle de la mémoire.
3	Elle s'identifia donc, avec une sorte de passion masochiste, à	**la vulnérable nudité de son visage**	sur lequel s'imprimaient toutes les traces de ses souffrances.
4	Je corrige donc mon diagnostic : « Le malade souffre de	**la déformation masochiste de sa mémoire. »**	En effet, il ne se souvient que des situations qui le rendent mécontent de lui-même.
5	… il ne comprend rien à	**la valeur de l'insignifiance.**	Voilà ma réponse à ta question sur le
6	En cela nous sommes tous égaux : Banaka, Bibi, moi et Goethe.	**L'irrésistible prolifération de la graphomanie**	parmi les hommes politiques, les chauffeurs de taxi, les parturientes, les amantes, les assassins, les voleurs […]
7	Les trois hommes nus, guidés par	**les injonctions muettes de son regard,**	se penchaient sur la petite et lui enlevaient le reste de ses vêtements.
8	Il imaginait	**l'histoire de son corps :**	il était perdu parmi des millions d'autres corps
9	Il faut comprendre	**le paradoxe mathématique de la nostalgie :**	elle est la plus puissante dans la première jeunesse quand le volume de la vie passée est tout à fait insignifiant.
10	Son bâillement n'avait ni commencement ni fin, c'était	**le bâillement infini de la mélodie wagnérienne :**	la bouche se fermait sans se clore tout à fait, elle s'ouvrait encore et encore, tandis que, à contretemps, les yeux aussi s'ouvraient et se fermaient.
11	[…] ils savaient fort bien qu'elles se répétaient avec	**la stupéfiante régularité de la probabilité statistique.**	Ils savaient que quelqu'un allait certainement

Figure 3. Les occurrences du schéma [Dét. + (Adj) + N1 + (Adj) + Dét. + N2]

Au niveau syntaxique, les occurrences de la figure 3 partagent la même construction, qui correspond à un **groupe nominal étendu** : un groupe nominal dans lequel le nom tête est modifié par un groupe prépositionnel modifieur du nom [Dét + N_1 + Prép + Dét + N_2]. Cette structure peut s'actualiser dans trois cas différents, selon que le nom tête est N1 ou N_2 (Riegel, M. *et al.*, p. 346-349) ; dans notre cas, le nom tête correspond à N_1.

D'un point de vue syntaxique, N_2 est toujours le complément du nom de N1, mais la relation entre les deux noms change. Il existe des relations de sujet à prédicat, où un verbe ou un adjectif est nominalisé (2, 4, 6, 7), par exemple la tournure « c'est l'exigence de la beauté » a été préféré à « la beauté exige que ...». La préposition *de* qui relie les deux substantifs peut indiquer d'autres types de relation, notamment la relation de possession au sens large, plus précisément de *propriété* ou de *qualité* (3, 8, 11), par exemple « la vulnérable nudité de son visage » a été choisi à la place de « son visage se caractérise par une nudité vulnérable ». Enfin, une relation attributive se forme entre les deux noms quand la préposition de instaure « un rapport de catégorisation discursive […] entre un nom à valeur générale classifiante et le référent particulier désigné par le nom tête complément » (Riegel, M. *et al.*, 1994, p. 348) : on ne dit pas que « la technique est froide et impersonnelle », mais on parle de « la froide impersonnalité de la technique » et ainsi de suite (1, 3, 9). De plus, le marqueur linguistique détecté par le *deep learning* (fig. 2) s'inscrit dans ce troisième cas : à « l'amour est agressif », « l'agressivité de l'amour » a été préféré. Dans de nombreux cas, cette relation attributive est également caractérisée par un langage figuré : par exemple, la nostalgie est comparée à un paradoxe mathématique. On

trouve également une tendance à la personnification (« la froide impersonnalité de la technique » ; « l'agressivité de l'amour ») et parfois à la concrétisation des concepts abstraits (« les flammes de l'extase »).
Le dénominateur commun de ces schémas est la présence d'une **prédication seconde** et d'un processus de **nominalisation**. En effet, un syntagme nominal a été préféré à la prédication première : par exemple, « son regard enjoignit aux trois hommes de » ou « de son regard, elle enjoignit aux trois hommes de » sont synthétisées par le GN « les injonctions muettes de son regard ». Dans le cas des relations de propriété, l'énoncé « elle se répétait avec la stupéfiante régularité de la probabilité statistique » a été choisi au lieu de « elle se répétait avec une régularité stupéfiante semblable à celle de la probabilité statistique ». Dans le troisième cas, la mémoire n'est pas déformée d'une façon masochiste, mais il est question de « la déformation masochiste de sa mémoire ». Ainsi, nous remarquons une tendance vers la nominalisation, le nom étant préféré au verbe ou à l'adjectif.
Dans certaines de ces occurrences, notamment dans les constructions attributives et dans certains des cas où il existe une relation de propriété entre les deux noms, on a affaire également à ce que Molinié définit comme la **métonymie de l'abstraction** (2011, p. 106). En effet, N1 correspond à une propriété de N2, propriété qui devient alors désignation. Par exemple, la structure « l'impersonnalité de la technique » permet d'accorder une plus grande importance à cette propriété, l'impersonnalité, qui aurait autrement été moins mise en valeur, si elle avait été présentée sous forme d'adjectif dans le syntagme « la technique impersonnelle » ou la phrase « la technique est impersonnelle ».

3.1.1. L'intention esthétique

Aller vers la nominalisation, vers le concept, permet ainsi de conserver les deux noms dans leur signification intégrale : avec « la technique est impersonnelle » il y aurait eu une déperdition du sens du concept signifié par l'adjectif. De plus, la prédication seconde et la nominalisation donnent lieu à un effet de synthèse : à une prédication première est préférée la force synthétique du groupe nominal étendu. Cette structure condense plusieurs significations relatives à un objet donné, en l'occurrence « la technique », de manière à évoquer dans l'esprit du lecteur une série de composants sémiques qui peuvent aider à mieux cerner sa réflexion. Il s'agit parfois de connotations qui prennent forme grâce à d'autres passages de son œuvre.
L'effet esthétique qui en résulte nous semble concorder avec sa poétique : ces schémas de phrases présentent un style concis qui, malgré sa précision et sa brièveté (ou grâce à elles) parvient à contenir – ou plutôt à donner à entrevoir – l'infinie richesse des connotations d'un thème ou d'une impression. De cette façon, ce choix assure ainsi une exploration du sens qui n'est jamais arrêtée (une

« irrésistible prolifération » de ses chemins) et qui se découvre continuellement stimulée.

Ce foisonnement du contenu sémantique de ces syntagmes nominaux rappelle le concept de variations sur un thème que Kundera lui-même théorise en référence à sa poétique : son œuvre est un grand roman contenant un certain nombre de *grands* thèmes qui sont développés sous forme de variations. Ces thèmes, « tels des fleuves, parcourent toute l'étendue de l'œuvre kundérienne sans se laisser interrompre par aucune frontière, ramassant sur leur passage et répandant dans chaque roman qu'ils irriguent un limon sémantique proprement inépuisable » (Ricard, 2003, p. 62). Ainsi, l'analyse existentielle de ces thèmes, ou, selon les termes de Kundera, l'exploration « d'une portion jusqu'alors inconnue de l'existence » est au cœur de sa poétique romanesque et c'est l'intention esthétique qui guide ses choix (Kundera, 2017, p. 707).

Il n'est donc pas inattendu que la significativité de cette structure ait été vérifiée par les approches statistiques et d'apprentissage profond, vu qu'elle réalise l'intention de saisir le foisonnement et la complexité des significations de plusieurs thèmes et situations existentielles, dont la composition sémique peut se trouver continuellement renouvelée et enrichie de nouveaux sèmes.

4. Conclusion

Nos hypothèses qualitatives, concernant la significativité d'une structure nominale particulière, ont été vérifiées grâce à l'apport à la fois de l'exploration statistique et de l'apprentissage profond ; plus précisément, le *deep learning* nous a permis d'en identifier des variantes et de relever, plus en général, la saillance des groupes nominaux.

Par conséquent, cette méthodologie intégrée nous a permis d'identifier et de définir un motif morphosyntaxique significatif de sa prose qui la distingue de celle de ses contemporains et de l'interpréter d'un point de vue linguistique et littéraire, de façon à déceler l'intention esthétique qui en sous-tend l'emploi. Cette démarche étant une nouvelle approche encore à explorer, nous visons à l'approfondir dans de futures études pour pouvoir détecter et examiner d'autres patrons morphosyntaxiques.

Bibliographie

Kalchbrenner N., Grefenstette E. et Blunsom P. (2014), « A convolutional neural network for modelling sentences », *52th Annual Meeting of the Association for Computational Linguistics*, p. 655–665.

Kundera M. (2017), *Œuvre I et II*, Paris, Gallimard, Coll. Bibliothèque de la Pléiade.

Legallois D. (2018), « Les motifs lexico-grammaticaux : une nouvelle approche en stylistique », in *Stylistique et méthode. Quels paliers de pertinence textuelle ?*, Lyon, Presses universitaires de Lyon.

Longrée D., Luong X. et Mellet S. (2008), « Les motifs : un outil pour la caractérisation

topologique des textes », in Serge Heiden, Bénédicte Pincemin (eds.), *JADT 2008*, Lyon, Presses de l'ENS.
Longree D., Mellet, S. (2013), « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraseologie de la langue au discours », *Langages* 189, p. 65–79.
Magri V. (2020), « Le deep learning comme défi pour identifier le style d'un écrivain : l'exemple de Jean Giono », *JADT 2020*, Jun 2020, Toulouse, France. ⟨hal-02936437⟩.
Mayaffre D. et Vanni L. (2021), *L'intelligence artificielle des textes*, Paris, Honoré Champion.
Molinié G. (2011), *Éléments de stylistique française*, Paris, PUF.
Ricard F. (2003), *Le Dernier Après-midi d'Agnès*, Paris, Gallimard.
Riegel, M., Pellat J.-C. et Rioul R. (1994), *Grammaire méthodique du français*, Paris, PUF, 2011.
Savoy J. (2016), « Estimating the probability of an authorship attribution », *JASIST*, 67(6), p. 1462–72.
Tuzzi A. Cortelazzo M. (eds. ; 2018), *Drawing Elena Ferrante's Profile*, Padova, PUP.
Vanni, L. et al. (2018), « Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis », Melbourne, *56th Annual Meeting of the Association for Computational Linguistics*, ⟨hal-01804310⟩.

Exploring Companies' communication strategies in the Covid-19 era

Alessandra Belfiore[1], Maria Spano[2],
Katerina Mandenaki[3], Walter Giordano[2]

[1]Università della Campania Luigi Vanvitelli – alessandra.belfiore@unicampania.it
[2]Università degli Studi di Napoli Federico II – maria.spano@unina.it walter.giordano@unina.it
[3]National and Kapodistrian University of Athens – mandenaki@yahoo.gr

Abstract
The unprecedented events of 2020 with the Covid 19 pandemic shocked economy. This study aims to identify the communication strategies adopted by companies to communicate their financial results to investors. We conducted a textual analysis of the letters to shareholders, downloaded from the annual reports of companies listed in the 2020 Standard & Poors's 500 Index. We performed a topic modeling and sentiment analysis procedure, to highlights the main topics and relative emotions. Then, we observed how these findings combined with different performance indicators. Our approach shows promising results in the effective analysis of corporate reports, as the financial narratives are of fundamental importance for determining market trends and influencing investment behavior.

Keywords: topic modeling, sentiment analysis, communication strategies, performance indices, annual reports

1. Introduction

The Covid 19 pandemic shocked not only ordinary life activities but also businesses and economies. The financial results of this difficult period were released in spring 2021 in companies' Annual Reports.

An annual report is a corporate document distributed to the shareholder that discloses the financial condition and operations of the company in the previous year. The key information, contained in the annual reports, is used by analysts to understand not only the current state but also the future direction of a company. Over the years these documents have become more and more powerful tools to convey not only financial performance but also corporate strategies, identity, and stance of business organizations (Garzone, 2004; Grove Ditlevsen, 2012). In

accounting and finance, the online availability of annual reports provides ample fodder for applying the technology (Loughran and McDonald, 2016).
This contribution aims to identify the communication strategies adopted, in times of generalized emergency, to transmit positive or negative financial results to stakeholders. We built a unique dataset with qualitative and quantitative data from the top 10% and the last 10% of the companies listed in the 2020 Standard & Poors's 500 Index. We performed a topic modeling allowed us to detect the thematic structure hidden in the letters to shareholders, detecting 10 recurring groups of related words. We performed a sentiment analysis to explore how emotion is used by the CEOs to communicate the financial results of the companies. Finally, we calculated the Pearson correlation coefficient to examine whether the affective and emotional values of the letters were correlated to the Total Return of the company, namely, if being a "loser" or a "winner" effectively increases or decreases the level of emotionality in the CEOs' discourse.

2. Data and methodology

2.1. Data Collection

The sample companies are extracted by the Standard and Poor's 500 index[1], (S&P 500 ®), as of Jan., 2022. The S&P 500 is a stock market index tracking the performance of large companies listed on stock exchanges in the United States. It is widely regarded as the best single gauge of large-cap U.S. equities. The index includes 500 leading companies and covers approximately 80% of available market capitalization. For the preliminary study we selected the top 10% of the best companies and the last 10% of the worst companies, listed in the index, with reference to the year 2020. The query counted 100 firms. For each company, we extracted the following quantitative and qualitative indicators. From the annual reports it was possible to recover the text of the letter to the shareholders from the CEO or the BOD, the headquarters, the sector and the industry in which the company operates. Instead, through the stock performance, set on Bloomberg®[2] closing prices – as of Dec. 31, 2019, and Dec. 31, 2020 – we define the annual Total Return. It includes dividends, capital gains, payments and distributions, and any other form of return and conveys a comprehensive view of the return on the investment. It measures how much the investment has grown over time and it is expressed as a percentage of the original investment.
We proceeded excluding firms for which complete information was not available for the years considered. Our final dataset is composed of 66 companies extracted by the Standard and Poor's 500 index (31 best and 35 worst).

[1] https://www.spglobal.com/spdji/en/indices/equity/sp-500
[2] https://www.bloomberg.com

2.2. Data Analysis

2.2.1. Topic Modeling

Topic modeling is basically a general term for a collection of methods and algorithms that uncover the hidden thematic structure in document collections by detecting recurring clusters of co-occurring words. There are several different algorithms for performing topic modeling, in this study we use the Latent Dirichlet allocation (LDA) (Blei et al., 2003). Topic modeling is mainly used to automatically and inductively structure the text to find recurrent patterns in the documents thus identifying the broader discursive fields. To perform LDA, texts have to be preprocessed following a set of steps well established in the literature on the statistical analysis of text (Wiedemann, 2016).

We used SpaCy NLP and Gensim, two libraries available in Phyton. SpaCy NLP library to tokenise texts, transform them into lowercase, and to lemmatize them in order to reduce the high variability of the vocabulary. Gensim library to identify the frequent bigrams in the corpus and append them to the list of tokens for the documents in which they appear.

Following a bag of words scheme, at the end of preprocessing we obtain a **documents x terms table T**(nxp), where the generic element t_{ij} is the frequency of the word j in the document i. Then, we remove from the vocabulary the words that appear less than 3 times and the words that appear in more than 25% of the documents. This was done to improve the quality of our model and speed up training.

For more robust results we implement the topic coherence metric, an in-built function of LDA topic modeling algorithm that estimates the number of topics by measuring the intrinsic coherence, namely the degree to which the words comprising the topic support each other (Röder et al., 2015). For that, we create many LDA models with different values for the number of topics and choose the one that provides the highest coherence value, namely a number that will generate meaningful and interpretable topics.

2.2.2. Emotional analysis on the documents

After determining the discursive and semantic fields of the proposed concepts we opt for a sentiment analysis of their discursive nature within the environment of our texts. Sentiment analysis is an important area of classification research where supervised and unsupervised methods are often used in order to classify text ordinally (from negative to positive, for example) rather than categorically. In this study we utilize emotion as a communication dimension to explore the way it is implemented within our letter texts to investigate the way affective vocabularies construct themes and representations, solidify conceptual spaces and emphasize the 'atmosphere' the speaker intends to create. For the purposes of this study, we opt for EMOLEX and two manually annotated lexicons: the

NRC Valence, Arousal, and Dominance Lexicon (NRC-VAD) (Mohammad 2018a) and the NRC Affect Intensity Lexicon (Mohammad, 2018b) that contains 10K English terms annotated for intensity (real-valued) scores corresponding to eight basic emotions (Plutchik 1980; Ekman 1992) although in our research we implemented the scores for anger, fear, joy and sadness. For the first step of our analysis, we measured the affectivity of all our texts. We investigate two affect-related features: the density of each affective dimension (valence, dominance and arousal) and the density of four emotions (joy, sadness, anger, fear), expressed in the body of the text. We compute the counts of affective words, each normalized by the total number of words in the CEO's discourses, and the total count of all the emotional words. We then calculated the Pearson correlation coefficient to examine whether the affective and emotional values of the letters were correlated to the Total Return of the company, namely, if being a “loser” or a “winner” effectively increases or decreases the level of emotionality in the CEOs’ discourse. Similarly, we also wanted to see if the size of the text (wordcount) would affect its emotionality and finally we checked if the emotional values were correlated to each other.

3. Findings

3.1. Topic Context

Topic modeling allows us to identify the main topics of the collection by detecting recurring clusters of co-occurring words. After running our coherence model, the results indicated that a number of 14 topics would offer the highest coherence score (each topic contributes a high percentage >0.6). We decided to show the first 10 words with the highest probability of belonging to the topic, for a better readability of the results (Figure 1).

Topics 0, 3, 5, 7,8, 11 and 12 can be interpreted as topics concerning problems related to lack of "liquidity", "bank", "loans", "loss" and costs related to "real_estate", "trade", "fiscal".

Indeed, topics 1, 9 and 13 are more closely related to the procedures faced during the pandemic. In this topic we find the topic concerning "therapy", "drug", "dose", "phase" of the pandemic and the "patient" that occupied hospitals during the Covid-19 era.

Topic 10 concerns the environmental improvement that occurred during the lockdown. In this topic we find words like "climate_change", "renewable", "natural", "pledge".

Finally, topics 2, 4, 6 and 10 concern remote working. These topics include "wireless", "computer", "fiber", "software, the distance of "colleagues" and the need for "news", "entertainment", "gaming".

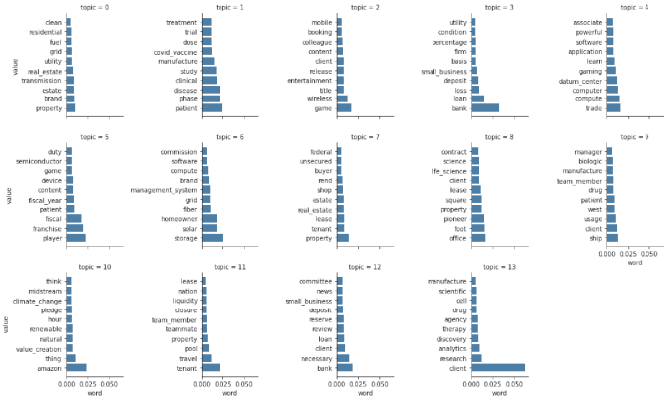
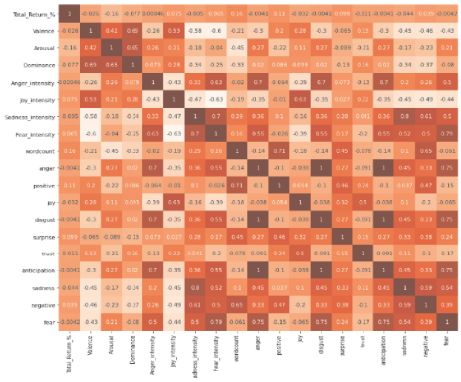


Figure 1. Best Topic

3.2. Relationship between communication strategy and performance

Figure 3 shows Pearson's correlation coefficient to examine whether the affective and emotional values of the letters correlated to each other and with the company's Total Return. There is the linear correlations between affective dimensions and emotions. It exists a very strong positive relationship between the Fear intensity score and the Sadness intensity score (r = 0.7) and the Anger intensity score (r = 0.63), while the relationship is strong but negative with the intensity of Joy (r = -0.63) and Valence (r = -0.60). Another very strong relationship persists between the score of Dominance and that of Valence (r = 0.69), and arousal (r = 0.65). Beyond the relationship that exists between the different affective and emotional values in the letters, there is no type of relationship between them with the total return of the company. This means that having a positive or negative performance did not significantly change the level of emotionality or affection in the CEOs' discourse. This is likely due to the fact that the Covid-19 pandemic was an unexpected shock. Therefore, as with all extraordinary events, no one knows how long the effects will be, regardless of whether they are positive or negative.



We then investigated the frequency of the emotional words on the basis of the companies' financial performance (Table 1). The aim is to find out if being a “loser” or a “winner” effectively increases or decreases the level of emotionality in the CEOs’ discourse.

Table 1. Emotional Words Use by Company Performance (Emolex)

	Fear		Anger		Sadness		Joy	
	Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**	**Word**	**Freq**
Profit Companies	pandemic	126	challenge	64	pandemic	126	cash	60
	challenge	64	cash	60	disease	39	share	53
	cash	60	demand	51	income	28	organization	49
	operation	49	disease	39	execution	24	proud	42
	disease	39	execution	24	margin	22	success	42
	change	33	wireless	19	theft	13	providing	37
	risk	26	powerful	19	console	13	create	35
	execution	24	theft	13	failure	11	progress	31
	advance	23	shock	11	support	11	achieve	31
	powerful	19	limits	11	limits	11	content	29
Loss Companies	pandemic	185	cash	74	pandemic	185	cash	74
	cash	74	challenge	67	debt	48	share	64
	challenge	67	demand	57	income	42	organization	53
	operation	66	loss	32	loss	32	create	47
	change	64	execution	19	lower	26	progress	42
	risk	34	rating	17	margin	21	income	42
	loss	32	restriction	13	execution	19	proud	41
	government	29	wireless	11	cases	18	success	40
	remains	24	injustice	10	rating	17	improvement	33
	advance	20	disruption	10	restriction	13	improve	32

For the first emotional dimension, that is fear, we note that while profitable firms focus on words such as "disease", "execution" and "powerful", the fear of loss-making firms is instead linked to "government", "loss" and "remains". The emotion of Anger for profitable firms concerns "theft", "shock" and "limits", while for loss-making firms we find words like "rating", "restriction", "injustice", "disruption". As for the dimension of Sadness, the emotional terms used by profitable companies such as "support", "bankruptcy", "closure", probably aimed at companies in the same sector that have faced difficulties during the pandemic period, are surprising. The emotion of sadness for loss-making companies, on the other hand, concerned "debts", "lower", "cases". Finally, the greatest difference between the terms is due to the emotion of joy. Among the most used terms for profitable firms, we find "proving", "achieve", "content" while for loss-making firms we have "income", "proud", "success", "improvement" and "improve".

4. Conclusions

We performed a textual analysis on letters to shareholders, extracted from the annual reports of a group of companies. Our strategy allowed us to investigate the main topics disclosed by CEO in the letters, but also their emotions and how the latter are related with their performance. The results of this preliminary study must be seen in the light of some limitations. Our results may be influenced by the selection of companies that we have retrieved from the Standard & Poors's 500 Index 2020. We could consider also other performance measures to have a more complete overview of the companies' financial results.

Future developments will be devoted to extend our study to all the other Companies listed the Standard & Poors's 500 Index, and to include also companies ranked by two other relevant stock market indices, namely the Nikkei and Eurostock.

References

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Ekman P. (1992). Are there basic emotions?.

Garzone G. (2004). Annual company reports and CEO's letters: Discoursal features and cultural markedness. *Intercultural aspects of specialized communication*, 14, 311-342.

Grove Ditlevsen M. (2012). Revealing corporate identities in annual reports. *Corporate Communications: An International Journal*, vol. (17): 379-403.

Loughran T. and McDonald B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), pp. 1187-1230.

Mohammad S. (2018). Obtaining reliable human ratings of valence, arousal and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 174-184.

Mohammad S. M. (2017). Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Plutchik R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pp. 3-33. Academic press.

Röder M., Both A. and Hinneburg A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399-408.

Wiedemann G. and Wiedemann. (2016). *Text mining for qualitative data analysis in the social sciences* (Vol. 1). Wiesbaden, Germany: Springer Fachmedien Wiesbaden.

AI and ML in accounting and finance: A bibliometric review

Alessandra Belfiore[1], Agostino Gnasso[2],
Corrado Cuccurullo[1], Massimo Aria[2]

[1]Università degli Studi della Campania Luigi Vanvitelli –
alessandra.belfiore@unicampania.it corrado.cuccurullo@unicampania.it
[2]Università degli Studi di Napoli Federico II –
agostino.gnasso@unina.it aria@unina.it

Abstract
Artificial intelligence (AI) and machine learning (ML) are two related technologies in accounting and finance studies. This study maps the conceptual structure of AI and ML research with the aim of contributing to a better understanding of this research stream. A bibliometric analysis of 3,836 documents on ai and ML retrieved from the Web of Science database is conducted. The analysis of descriptive performance indicators identifies the main traits of the scientific debate about AI and ML in terms of publications, productive countries and sources. To map the conceptual structure of the dataset, the study performs a thematic evolution. The results highlight the growing academic interest in the research topic, especially in the past few years. The results of this study may provide scholars with a better understanding of AI and ML research in accounting and finance. This paper contributes to the field by providing an examination of the current state of the art of AI e ML research and identifying possible future research directions.

Keywords: Accounting, Finance, Artificial Intelligence, Machine Learning, Conceptual structure, Journal ranking

1. Introduction

The growing success of artificial intelligence (AI) and machine learning (ML) in the application and research domains of accounting and finance is revolutionizing industries and corporations (Li and Tang, 2020). Hedge funds, retail banks, and FinTech, supported by massive growth in computing power and storage, are investing heavily in the acquisition of skills in data science and ML (Wall, 2018). Furthermore, together with the continuous search for innovation with the use of big data, there is an ever increasing need to observe regulatory systems. For example global financial crisis in 2007 led to major structural changes in the financial industry, pushing tight regulation on the topic of big data. De Prado (2019) argues that impede to use classical econometric linear

models are unable to predict due to different types of data, such as using data such as voice recordings, news articles, social media posts media such as Tweets, photos and satellite images, called alternative data. Hence, ML models allow to provide computational power and flexibility to run complex models in high-dimensional data environments. Moreover, the scientific community searching for more accuracy and interpretability of these ML models, are starting to use models that exploit the internal relationships between features in order to perform interpretative analyses, predictive analyses and casual inference (Aria, Gnasso, Cuccurullo, 2021).

This study aims to highlight how AI technology and ML models have influenced the accounting and finance industries in the last decade. The contribution to the existing literature can be identified from several points of view. In particular, this study examines the related AI-Accounting and Finance literature, published between 2010 and 2021, and summarizes its focused areas, allowing to highlight the trend and nature of the contents found in the current literature. Secondly, this work shows the current application of AI and ML in the fields of accounting and finance, which aims to highlight the relevant areas, and can certainly serve as a reference point for future studies in this domain.

2. Materials and Methods

2.1. Data collection

To retrieve the bibliometric data, we used the Web of Science (WoS), querying SCI-EXPANDED (Science Citation Index Expanded) and SSCI (Social Science Citation Index) databases. These two databases appear to be the most used in the social sciences (Zupic, Čater, 2015) and have standardized fields that require little data cleaning. We used the following search terms: "Artificial Intelligence", "Machine Learning", "Deep Learning", "Data Science" in the title, abstract, and keywords fields over the period 1985-2022, focusing only on English documents. Few articles had been published before 2010, while in 2022 publications are still in progress when the data was downloaded. So, we focused on the 2010- 2021 period. Also, there were various document types, but we decided to only consider the articles, proceedings paper, review. Finally, we focused only on two subjact area: "business", "finance" and "management". This query returned 3,836 publications.

2.2. Data analysis

We analyzed the whole dataset with the Bibliometrix R package (Aria, Cuccurullo, 2017). It is a unique tool according to a logical bibliometric workflow and incorporates a wide variety of different analyses. This tool is the only one that allows you to control and follow the entire bibliometric analysis workflow from importing bibliographic data from various databases, including the Clarivate Analytics Web of Science, to performing bibliometric analysis and

data visualization. It is open-source and returns statistical analysis that other bibliometric tools do not provide.

After some descriptive statistics, we extracted the conceptual knowledge structure of AI and ML research in accounting and finance. This structure is used to understand the topics covered by a research field to define the most important and recent issues (the so-called research front) and to study the evolution of concepts over time (Börner et al., 2003). The methodological foundation of the conceptual structure is the idea that the co-occurrence of key terms describes the content of the documents in a dataset (Callon et al., 1991, Cobo et al., 2011). We used the KeywordsPlus to identify thematic clusters.

By applying the Louvain clustering algorithm on the co-word network, we highlighted the different themes of a given domain. This thematic network is drawn on a two-dimensional matrix, where the axes are function of the Callon centrality and the Callon density of the thematic map. The centrality can be read as the relevance of the theme in the whole research field. Density is intended as a measure of the degree of development of the theme.

The two-dimensional matrix defines four quadrants related to different themes (Callon et al., 1991; Coulter et al., 1998; He, 1999; Cahlik, 2000; Aria et al., 2020): (i) motor themes (upper right quadrant) which are very relevant and very well developed; (ii) basic themes (lower right quadrant) that are very topical but not well developed; (iii) emerging or declining themes (lower left quadrant) not very relevant and developed; (iv) niche themes (upper left quadrant) not very relevant but very well developed.

Each theme represented in the bi-dimensional matrix is a co-word network cluster. The names in the bubble are the words with the highest occurrence value. The size of the bubble is proportional to the occurrences of the words in the cluster.

For a longitudinal analysis, we articulated the time span into 3 time slices. Looking at the distribution of publications per year, we have decided to divide our collection into 3 equal time spans (2010-2013; 2014-2017; 2018-2021), with 2 cutting points (2013 and 2017).

3. Findings

3.1. Descriptive findings

Table 1 shows significant information about our collection. The number of documents on AI and ML is growing exponentially. In the period from 2010 to 2021, the annual growth rate is 16.71%. This exponential growth rate indicates that academic production is intensifying. This is probably due to the fact that the topic of AI and ML in accounting and finance is global. The analysis of national scientific production showed that the number of documents in which at least one author comes from a particular country varies from 1 for Iraq, Montenegro, Philippines to 3,730 for the USA (Figure 1).The most productive countries for

publications related to AI and ML in accounting and finance are the United States of America (USA) (n=3.730; 27%), followed by China (n=1.677; 12%), the United Kingdom (UK) (n=1,247; 9%), Germany (n=795; 6%).
The 3,836 documents were written by 9,080 authors on 299 different journals that are all non-leading journals (ABS 2018 ranks < 3). This reflects the ABS 2018 ranking in which 70% (62/89) of accounting journals and 66% (73/110) of finance journals are ranked < 3.
As shown in Figure 2, the most relevant source is Technological Forecasting and Social Change with 258 publications (7%). Approximately, 4% of the documents are published by European Journal of Operational Research (n=165) and Research Policy (n=148). Additional sources for AI and ML include the following journals: Journal of Business Research (n=112; 3%), Science and Public Policy (n=84; 2%), Journal of Technology Transfer (n=70; 2%), and International Journal of Forecasting (n=67; 2%).
All the publications are associated with 11,082 authors' keywords and 5,915 keywords assigned by the WOS machine learning algorithm.

Table 1. Main information about data

MAIN INFORMATION ABOUT DATA	
Timespan	2010:2021
Sources (Journals, Books, etc)	299
Documents	3836
Average citations per documents	22,19
Average citations per year per doc	3,577
References	181027
DOCUMENT TYPES	
article	3578
article; book chapter	7
article; proceedings paper	45
review	206
DOCUMENT CONTENTS	
Keywords Plus (ID)	5915
Author's Keywords (DE)	11082
AUTHORS	
Authors	9080
Author Appearances	11366
Authors of single-authored docs	467
AUTHORS COLLABORATION	
Single-authored docs	486
Documents per Author	0,422
Co-Authors per Doc	2,96
International co-authorships %	38,35

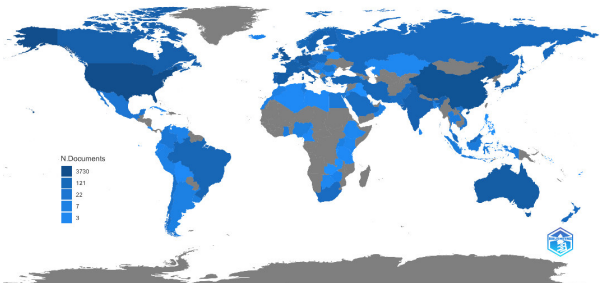


Figure 1. Country scientific production

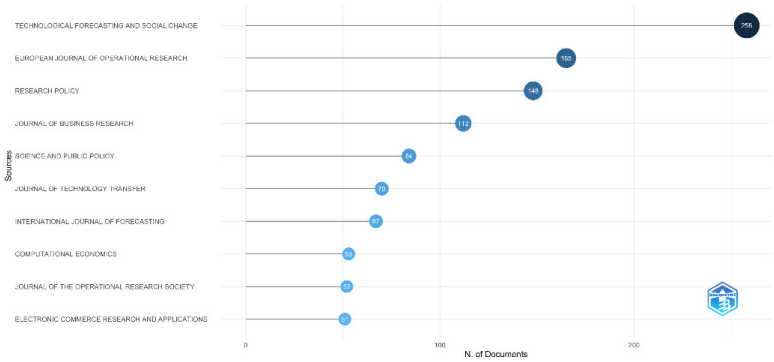


Figure 2. Most relevant sources

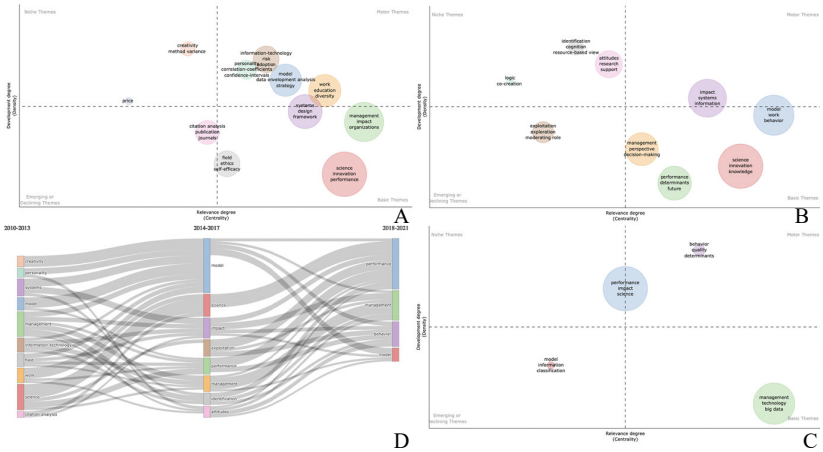


Figure 3. A, time slice 2010-2013; B, time slice 2014-2017; and C, time slice 2018-202; D, Thematic evolution of Keywords Plus in field of research on AI and ML (2010-2021).

3.2. Conceptual structures evolution

Figure 3 reports the results of the analysis of the thematic evolution of publications in the three periods considered (2010-13; 2014-17; 2018-21). Looking at the Callon centrality and density, we analyze the motor themes, the basic themes, the emerging or declining themes, and the niche themes. Comparing the thematic map of the three periods considered, we analyze and trace the evolution of the themes (in terms of trajectory on the theme). In the

first period 2010-13, the research identified 10 thematic areas (Figure 3B): "information-technology", "strategy", "education", "personality" as a motor theme, "systems", "management", "innovation" and "ethics" as a base theme, "citation analysis" as an emerging theme and "creativity" as a niche theme. In the second period 2014-17 (Figure 3C), all the motor themes become basic or niche themes. Finally, in the third period 2018-21, we have a concentration in 4 thematic areas (Figure 3D). The most relevant and developed topics in the research on AI and ML in accounting and finance is "quality" and “impact” themes. Instead, "management" is a basic theme, while "information" as an emerging or declining theme.

4. Conclusion

This article examined the existing literature on AI and ML to study the conceptual framework of research in accounting and finance. We performed a bibliometric analysis divided into two steps: first, we have made a descriptive analysis; second, we have identified the conceptual knowledge structure. Our findings highlight a growing and global academic interest in this research topic. Recommendations on what potential journals could be selected for publication. We conclude that no studies have been published in top journals. The analysis of the thematic evolution has allowed us to understand the research themes and/or thematic changes that have accompanied the growth of scientific production on AI and ML in accounting and finance. Keyword Plus analysis shows how AI and ML research in accounting and finance is focusing on quality, operations, impact and performance. Our study has both theoretical and practical implications. Firstly, this study contributes to the existing literature on the topic of AI and ML in accounting and finance. This analysis provides a starting point for scholars interested in studying this research topic in several ways. The study presents the longitudinal evolution of AI and ML research, revealing the most important themes related to AI and ML of each period. The conceptual structure shows that criticalities have already been developed, while theoretical and empirical areas remain to be explored. Secoondly, our research highlights how the term “impact” has become relevant in the conceptual map. This may also present limits from which to define potential directions for future research.

This study focused on international articles, paper proceedings, and peer-reviewed publications only in journals indexed on WoS. As a result, different types of data on the subject (such as patents) were not included in our analysis. We have limited our investigation to the Plus keywords only. It would be interesting to develop further conceptual analyzes (such as LDA).

References

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

Aria M., Cuccurullo C. and Gnasso A. (2021). A comparison among interpretative proposals for random forests. Machine Learning with Applications, 6, 100094.

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.

Börner K., Chen C. and Boyack K. W. (2003). Visualizing knowledge domains. Annual review of information science and technology, 37, 179– 255.

Cahlik T. (2000). Comparison of the maps of science. *Scientometrics*, 49, 373–387.

Callon M., Courtial J.-P. and Laville F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22, 155–205.

Cobo M. J., L´opez-Herrera A. G., Herrera-Viedma E. and Herrera F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for information Science and Technology*, 62, 1382–1402.

Coulter N., Monarch I. and Konda S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49, 1206–1223.

He Q. (1999). Knowledge discovery through co-word analysis.

Harvey C., Kelly A., Morris H. and M. Rowlinson (2010), *Academic Journal Quality Guide*, Version 4, London: The Association of Business Schools.

Li X. and Tang P. (2020). Stock index prediction based on wavelet transform and fcd-mlgru. *Journal of Forecasting*, 39, 1229–1237.

L´opez de Prado M. (2019). Beyond econometrics: A roadmap towards financial machine learning. Available at SSRN 3365282, .

Wall L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, 100, 55–63.

Zupic I. and Cater T. (2015). Bibliometric methods in management and organization. Organizational research methods, 18, 429–472.

Complex networks and natural language processing to unfold the social enterprises environment

Loredana Bellantuono

Università degli Studi di Bari Aldo Moro –
loredana.bellantuono@uniba.it

Abstract
The study of socio-economic systems through mathematical methods is an established technique of both description and prediction. The increasing availability of data and computing resources has led to a new approach based on the science of complex systems, which helps grasping and interpreting information hidden in large amounts of data, that cannot emerge through conventional statistical methods (Zeng et al., 2017). The proposed research work aims at measuring the territorial impact of social economy projects, through the implementation of quantitative methods that allow to extract relevant information from large amounts of heterogeneous data (Bellantuono et al., 2020b, 2022). The research is part of a project promoted by the Invitalia Operational Think Thank (OTT), to transform the "Italia Economia Sociale" initiative, funded by the Italian Ministry of Economic Development, into a pilot intervention of Social Smart Land (Bellantuono et al., 2020a).
We model the national system of investments in social inclusion projects combining Complex Network methods (CN) with Natural Language Processing (NLP), focusing on a specific technique known as Latent Semantic Analysis (LSA), that maps relationships between a set of documents and the words they contain, through the creation of a set of concepts (word space) (Dumais, 2004). By implementing this CN+NLP approach, we examine the public database of the Open Coesione portal, containing information on over 340000 projects related to the social inclusion thematic area, funded under the 2007-2013 and 2014-2020 cohesion policy campaigns. The organizations involved in such projects are social enterprises and other legal entities with characteristics similar to those required for participation in the call "Italia Economia Sociale". The ecosystem of social economy fundings is modelled as a complex network, with nodes representing social enterprise projects. A link between two nodes is present if a similarity index, based on semantic affinity of project titles and abstracts, is larger than a fixed threshold. The network is analyzed with an unsupervised community detection algorithm (Newman, 2006), unveiling the presence of 8 prominent groups, overall collecting 29.13% of the network, formed by nodes that tend to connect among each other rather than to external nodes. To investigate the common traits of nodes in each community, we use a wordcloud, namely a visual representation of the verbal content of a document

in which the size of terms with a non-trivial semantic meaning increases with their frequency.
The described CN+NLP method represents a fully data-driven approach, offering a unique cutting-edge tool to map unexpected similarities in the social enterprise environment, which can be efficiently exported to different business areas.

Keywords: Complex systems; Complex Networks; Natural Language Processing; Community detection; Wordclouds; Social Economy; Social Smart Land.

References

Bellantuono L. and Bellotti R. (2020a). Premessa: Modello di rete complessa per la ricerca di indicatori di impatto degli investimenti in economia sociale. *DIAMO FUTURO ALL'ECONOMIA SOCIALE – Il Terzo Settore e i progetti per l'inclusione sociale finanziati nell'ambito delle politiche di coesione 2007-2020, Working Paper Invitalia*. https://www.invitalia.it/cosa-facciamo/rafforziamo-le-imprese/italia-economia-sociale/invitalia-social-smart-land/documenti

Bellantuono L., Monaco A., Tangaro S., Amoroso N., Aquaro V. and Bellotti R. (2020b). An equity-oriented rethink of global rankings with complex networks mapping development. *Scientific Reports, 10*, 18046. https://doi.org/10.1038/s41598-020-74964-3

Bellantuono L., Monaco A., Amoroso N., Aquaro V., Bardoscia M., Demarinis Loiotile A., Lombardi A., Tangaro S. and Bellotti, R. (2022). Territorial bias in university rankings: a complex network approach. *Scientific Reports*, 12, 4995. https://doi.org/10.1038/s41598-022-08859-w

Dumais S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38, 188. https://doi.org/10.1002/aris.1440380105

Newman M. E. J. (2006). Modularity and community structure in networks. *PNAS, 103(23)*, 8577. https://doi.org/10.1073/pnas.0601602103

Zeng A., Shen Z., Zhou J., Wu J., Fan Y., Wang Y. and Stanley H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, 714-715, 1. https://doi.org/10.1016/j.physrep.2017.10.001

Fake news spreading and sentiment of Italians during the first COVID-19 lockdown

Pietro Belloni[1], Margherita Silan[1], Giulia Cuman[2]

[1]University of Padua, Italy– pietro.belloni.1@phd.unipd.it
silan@stat.unipd.it.it
[2]Research Ethics Committee, University Hospital of Padua, Italy–
giulia.cuman@gmail.com

Abstract
The last optional question of the SEBCOV online survey is open-ended and asks for additional comments: it received a particularly high response rate. In these answers, it is possible to read personal experiences about the first lockdown (March, April 2020) given spontaneously, without the researcher's influence. We performed a sentiment analysis dividing the textual answers into three groups: positive, negative, or neutral content. We have also noticed that several answers contained fake news, thus we flagged them differently. Then, to evaluate the joint effect of other survey variables on the sentiment and on the fake news spreading, we fit two specific models: a Bayesian regression model to explain the respondents' sentiment; and, a logistic model to deepen the fake news spreading phenomenon. We found that respondents who experienced an income loss during the lockdown more frequently reported a fake news as true in their comments at the end of the questionnaire.
Our analysis provides a novel insight about Italians' hurdles during lockdown, highlighting categories that were more affected, and contributes to understanding the determinants of fake news spreading during the pandemic.

Keywords: Sentiment analysis; Fake news; COVID-19; Survey; Open-ended questions.

1. Introduction

In December 2019, a novel coronavirus, SARS-CoV2 (Severe Acute Respiratory Syndrome Coronavirus 2), which causes the respiratory disease COVID-19, was initially discovered in Wuhan, China (Yan *et al.*, 2020). At the beginning of 2020, SARS-CoV2 started to spread rapidly all over the world and was subsequently declared a "Public Health Emergency of International Concern" by the World Health Organisation (30th January 2020). On 11th March 2020 this novel coronavirus had spread so severely that the WHO mutated its status to a "Global Pandemic" (Guo *et al.*, 2020).

There is a multitude of studies that tackle COVID-19 and nonpharmaceutical interventions (NPIs) - such as social distancing, isolation of infected individuals, quarantine, travel restrictions, and use of personal protective measures - from different perspectives: the SEBCOV study (Panngum *et al.*, 2020) specifically aims to investigate the impact of NPIs on the population from a societal and ethical point of view. It involves five different countries: Italy, Slovenia, Malaysia, Thailand, and the United Kingdom. In this work, we focus only on Italian data: in Italy, the study began at the end of the first national lockdown that started on 11th March 2020 and ended on 4th May 2020.
The analysed data come from the SEBCOV survey, which has 36 questions concerning social, ethical, and behavioural aspects of the COVID-19 pandemic. The questionnaire is composed of five sections: (1) socio-demographic information; (2) income, occupation status, and economic impact of COVID-19; (3) preferences and perceptions regarding COVID-19-related communication and occurrence of fake news; (4) perceived level of knowledge about COVID-19, the use of non-pharmaceutical interventions and behavioural changes; (5) concerns and coping strategies related to restrictions (Panngum *et al*., 2020). However, the focus of the analysis is not on the topics addressed by the questionnaire but on the last open-ended question *"Is there anything else that you would like to tell us?"*. Indeed, the response rate for this optional question was surprisingly high (34%): we exploit the information contained in the answers to that last question to get a valuable, spontaneous, and original insight about Italians' reaction to the first lockdown.

2. Data

The study participants were adults who have provided informed consent, residing in Italy and able to use a computer or a smartphone. The survey has been administered using the same questionnaire in two different ways. From 21st April to 4th May 2020 the survey was promoted through social networks (Instagram and Facebook) following a quota sampling to get a balanced sample with respect to gender, age, and geographical region. In the period chosen for the survey, Italian citizens were living the last days of the first period of lockdown, ended on 4th May 2020. On the other hand, from 1st May to 30th June the same questionnaire was publicized and promoted on Facebook on a page dedicated to the SEBCOV research.
The majority of answers to the optional question came from the first shot (352 out of 575); however, differences in the compositions of respondents to the two shots of the survey are not relevant in the analysis of the last optional question. Thus, in our analyses, only respondents to the last optional question are considered regardless of the survey shot they took part in.

3. Methods

We performed a sentiment analysis dividing the textual answers into three sentiment groups based on the content: positive, negative, or neutral.

First, an attempt to extract the sentiment from the text using the R library *TextWiller* (Sciandra *et al*., 2016) was made. *TextWiller* is an R package specifically developed for text mining and sentiment analysis in the Italian language: it allows to classify the answers to the open question of the survey as "positive", "neutral" and "negative", but the overall result was clearly poor. Then, an attempt with the R library *SentimentAnalysis* (Feuerriegel and Proellochs, 2019) was made. The library processes only English text, so the Italian text has been automatically translated with the machine learning-based translation system DeepL (*https://www.deepl.com/translator*) prior to running the sentiment analysis routine. Since this second attempt was also not satisfactory, the sentiment classification was conducted manually: two authors have independently read the answers and assessed their sentiment. In 82% of cases an agreement was found in the classification; the remaining cases have been discussed on a case-by-case basis, so a final sentiment has been established for all answers. This sentiment classification can be considered a gold standard and allows to confirm the poor performance of the two automated techniques measuring the agreement between the previous classification and the gold standard. The former has a Cohen's kappa of 0.25 (95% CI: 0.19 - 0.31) and the latter has a Cohen's kappa of 0.19 (0.14 - 0.24). The difficulties of automated sentiment classification methods arise because the available text is much more complex than the text usually analysed via sentiment analysis (i.e., text from social media or review websites): in many cases one sentence expresses more than one sentiment or presents misspells, colloquial jargon or syntax inaccuracies.

Besides establishing the sentiment, the answers to the open questions have been independently classified according to the presence of fake news. An answer was classified as fake news if stating one or more facts in clear contrast with the scientific truth.

To evaluate the joint effect of the survey variables on the sentiment, a Bayesian regression model has been fitted using Markov chain Monte Carlo: the response variable is the three-level sentiment, regression features have been selected from the survey variables using a stepwise procedure comparing different models via leave-one-out cross-validation (Vehtari, Gelman and Gabry, 2017). The model has two intercepts (one for neutral coefficients and one for positive coefficients, the negative class is treated as baseline) and on both a non-standardized t-student prior distribution $t_3(0, 2.5)$ is assumed (Jackman, 2009). In this setting, due to the presence of multiple intercepts, choosing a Bayesian model ensures a clearer analysis of the effects of the features on the sentiment given by the probabilistic representation of the posterior distribution. The model goodness-of-fit has been

evaluated using a graphical posterior predictive checking (Gelman *et al*., 2013). To evaluate the joint effect of the survey variables on the fake news spreading, a logistic model has been fitted using maximum likelihood. The response variable is the two-level presence/absence of fake news in a text, and the regression features have been selected from the survey variables using a stepwise selection based on AIC. The model goodness-of-fit has been evaluated using the likelihood-ratio test and the AUC calculated on the predicted data.

4. Results

A total of 575 respondents answered to the optional question of the survey in two shots; 77 of which have been discharged due to not meaningful text (e.g.: blank text, not relevant text…). The average answer length was 25.43 words (with a standard deviation of 32.28) and the most frequent words and couples of words (bigrams) are represented in Figure 1. Many recurrent words and bigrams are expected, such as *virus* (word written 33 times), *before* (word repeated

Figure 1. Most recurring words (left) and bigrams (right), font size is proportional with the words count (drawn up with https://www.wordclouds.com/).

33 times, usually part of sentences like before the pandemic my economic condition was better), *I cannot* (bigram used 12 times, usually part of sentences like I cannot go outside), *social distancing* (bigram written 8 times), *stay at home* (bigram used 6 times), *other region* (bigram repeated 6 times, usually part of sentences like I cannot travel to another region). Many others reflect criticism and dissatisfaction about restriction policies: this is the reason why words and bigrams such as *Italian government* (bigram used 8 times), *layoffs* (bigram written 7 times), *job* (word repeated 46 times, usually part of sentences like due to lockdown restrictions, I lost my job) or *government* (word written 47 times) are frequent.

In 278 (56%) answers the text was associated with negative sentiment, in 181 (36%) answers the text was associated with neutral sentiment, and only in 39 (8%) answers the text was associated with positive sentiment. One (or more)

fake news was present in 84 (17%) of the texts.
The coefficients of the model are shown in Figure 2: each point represents the predicted probability (with the associated interval) of a sentiment in the corresponding population subgroup. Once adjusted for all other variables, a higher probability to express a negative sentiment among men than among women is observed. In general, young respondents are more pessimistic than older individuals; indeed, it is more likely to observe a positive sentiment among old respondents (65 years old or more) than among those younger than 45 years old. Thus, according to answers to the survey, young people are those who have suffered the most; indeed, 70% of them had mental health concerns (for example: boredom, loneliness, anxiety, depression), and 74% suffered from the lack of social contact (with friends or with the non-co-resident partner). Furthermore, young people are strongly concerned about their working condition. Another variable that is fundamental to explain the sentiment expressed by respondents is the geographical region where they live: respondents living in the North of Italy have a higher probability to express a negative sentiment than those living in the South of Italy. Indeed, the COVID-19 pandemic started in Italy with two outbreaks; one in Lombardy and one in Veneto (Odone *et al*., 2020) and cases were higher in the North of Italy than in the South during the time of the survey. Moreover, respondents who also had an income loss caused by the pandemic are

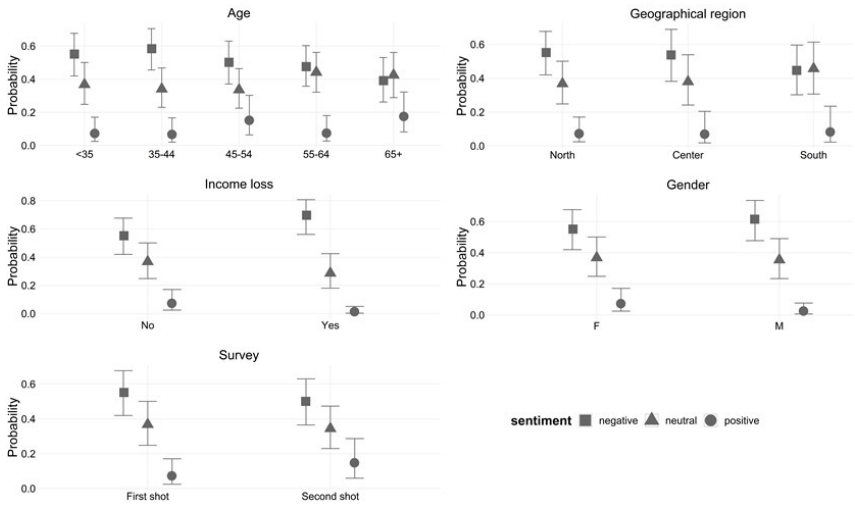


Figure 2. Estimated sentiment probabilities for the classes of the selected variables (intercepts not shown).

more likely to write a negative sentence in the last optional question of the survey. The lockdown of the COVID-19 pandemic has been having heavy effects on the Italian labour market and on most of the production activities, which have

reverberated in the entire economy (Fana, Torrejón Pérez and Fernández-Macías, 2020).

Table 1. Coefficients of the model for fake news presence (intercept not shown) expressed as odds ratios (class vs baseline).

Predictors	Odds Ratios (95% CI)	p-value
Gender (*male* vs *female*)	1.85 (1.12-3.07)	**0.02**
Income loss during lockdown (*yes* vs *no*)	1.76 (1.06-2.93)	**0.03**
Self-judgment of the ability to detect fake news (*high* vs *low*)	1.42 (0.97-2.14)	0.08
Gets informed using traditional media (*no* vs *yes*)	4.10 (2.10-7.97)	**<0.01**
Gets informed using institutional media (*no* vs *yes*)	2.19 (1.26-3.79)	**<0.01**
Survey shot (*second* vs *first*)	0.69 (0.37-1.23)	0.22
Observations: 498		
AUC: 0.70		

A logistic generalized linear model has been fitted to evaluate the presence of fake news in a text, the coefficients are shown in the form of odds-ratios in Table 4. Men are 85% more likely than women to write fake news in text and respondents who experienced an income loss during lockdown are 76% more likely to do it as well. Respondents who declare to get informed by traditional and institutional media write fake news in their comments less frequently. The aspect that looks contradictory is that those who self-judge able to detect fake news are also those who more easily include fake news in the final optional question. The 95% of these respondents, who self-judge able to detect fake news but report fake news in the optional question at the end of the questionnaire, get informed online and only the 72% exploits traditional media, which is a much lower percentage than in the rest of the population (93%).

5. Conclusion

The final question of an elaborate international survey unexpectedly collected much nonstructured data and proved to be a source of valuable information about the social struggle of Italians through the first COVID-19 lockdown. In conclusion, this study shows some interesting results about the sentiment freely expressed in the open-ended question. The fake news spreading is also taken into account and analyzed.

Both sentiment and propensity to write fake news are influenced by the gender and the income loss of the respondent. Moreover, age and residence area have an influence on the sentiment as well as the self-judgment of the ability to detect fake news and the media diet of the respondent heavily influence the propensity to write fake news.

References

Fana M., Torrejón Pérez S. and Fernández-Macías E. (2020) Employment impact of Covid-19 crisis: from short term effects to long terms prospects, *Journal of Industrial and Business Economics*, 47(3), pp. 391–410. doi: 10.1007/s40812-020-00168-5.

Feuerriegel S. and Proellochs N. (2019) SentimentAnalysis: Dictionary-Based Sentiment Analysis.

Gelman A. *et al*. (2013) *Bayesian data analysis*. CRC press.

Guo G. *et al*. (2020) New Insights of Emerging SARS-CoV-2: Epidemiology, Etiology, Clinical Features, Clinical Treatment, and Prevention, *Frontiers in Cell and Developmental Biology*, 8, p. 410. doi: 10.3389/fcell.2020.00410.

Jackman S. (2009) *Bayesian analysis for the social sciences*. John Wiley & Sons.

Odone A. *et al*. (2020) COVID-19 deaths in Lombardy, Italy: data in context, *The Lancet Public Health*, 5(6), p. e310. doi: https://doi.org/10.1016/S2468-2667(20)30099-

Pan-ngum W. *et al*. (2020) Social, ethical and behavioural aspects of COVID-19, *Wellcome Open Research*, 5(90). doi: 10.12688/wellcomeopenres.15813.2.

Sciandra A. *et al*. (2016) TextWiller: Collection of functions for text mining, specially devoted to the italian language.

Vehtari A., Gelman A. and Gabry J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, 27(5), pp. 1413–1432. doi: 10.1007/s11222-016-9696-4.

Yan Y. *et al*. (2020) The First 75 Days of Novel Coronavirus (SARS-CoV-2) Outbreak: Recent Advances, Prevention, and Treatment, *International Journal of Environmental Research and Public Health*, 17(7).

A Preparatory Corpus-Based Study of Swiss Touristic Brochures. Euphoric markers and Phraseology

Marion Bendinelli

University of Bourgogne-Franche-Comté –
marion.bendinelli@univ-fcomte.fr

Abstract
Tourism discourse is a hybrid discourse, articulating informative, advertising, epidictic and procedural features. Various genres exist: we study the specific case of brochures. The paper consists in a preliminatory study of both a parallel and comparable corpus of Swiss tourism brochures in French and Italian. With the softwares TXM and mkAlign, we outline two different ways to consider the brochures' informative and persuasive functions by looking at the use of euphoric markers and phraseological segments of quantity and intensity. The study also invites to think about the translation processes adopted in the brochures.

Keywords: Tourism discourse, Multilingual brochures, Parallel and comparable corpus, Euphoric markers, Phraseology

1. Introduction

Tourism discourse is a specialized discourse: it displays the fundamental features of general language and some specific lexical, syntactic and semantic features (Maci, 2020). These specificities are linked to the special communication contract (Charaudeau, 2009) that bonds the publicist to the professional s/he answers to, the competitors and the audience. Tourism discourse informs the audience on the characteristics of a given product (a destination, a monument...) and draws a list of practical recommendations; it affects the audience perceptions and turns the reader into a potential tourist to have him/her travel. To reach this goal, touristic texts rely on superlatives, on attractive visuals and on the community values in order to depict an idyllic and most appreciated land. For these reasons, tourism discourse is intrinsically hybrid, articulating features of informative, advertising, epidictic and procedural discourses. Various genres instantiate tourism discourse. Nigro (2006) recognizes five traditional categories – leaflets, brochures, posters, travelogues and travel guides – to which Maci (2020) adds more hybrid and less institutionalized forms. Our work focuses on the case of brochures, defined as a means of communication addressed to non-

professionals and read through prior or during the trip. According to Maci (2020), brochures are highly persuasive to sell tourists products. They must be attractive and efficient, highlighting the geographic characteristics of a destination, the available infrastructures and activities, the existing offers. Variable in size, brochures look like small books: a cover, a title, a table of content, subheadings and/or visuals dividing portions of texts (Huigen, 2004 in Carron, 2015). In brochures, the speaker is generally unidentified, his/her voice mostly erased; the addressee is made explicit with the use of the second person pronoun but not always identified in his/her social characteristics, except in the case of targeted promotional texts (single persons, families…). Previous studies have outlined two main discourse characteristics. Brochures (1) are intrinsically expressive: they make an intensive use of emotions, euphoric markers and stylistic figures; (2) break with reality and everyday life through the isotopies of magic and fantasy, of time and space travel in order to depict an atemporal world. The present paper describes different analytical entries to investigate brochures' expressive function, in particular euphoric terms and phraseology. It adopts a multilingual point of view: the next section describes our exploratory corpus.

2. Swiss Touristic Brochures: a Parallel and Comparable Corpus

Swiss tourism offices publish a large range of touristic materials: catalogues, leaflets, brochures, small and large maps, postal cards and goodies among others. While leaflets are generally edited in three languages (French, English and German; Italian, Spanish and Portuguese), brochures are intrinsically monolingual (French, English, German or Italian). Some of them may be downloaded from the office website. We have collected 17 brochures and 15 leaflets and worked so far on four brochures alone: brochure *Grand Tour* in its French and Italian versions, the brochure named *N°01* in Italian and the brochure *N°03* in French which are both structured by the same table of contents (Mountain, Water, Housing/Environment, Inhabitants, Concierge Services). Hence, our corpus is bilingual, parallel (both *Grand Tour* brochures are strictly equivalent yet in two different languages) and comparable (brochures *N°01* and *N°03* deal with the same topics, arranged in the same way, yet with different contents and in two languages). We digitized and ocerized (with ABBYY Finereader 9.0) these 4 brochures, then established them in XmL-TEI compliant language. The French part of the corpus (FRTOURISM) gathers 21.098 word forms and the Italian part (ITTOURISM) 21.865. Texts were eventually lemmatized with TreeTagger, while implemented in the software TXM (Heiden, Magué, Pincemin, 2010). Also, we uploaded the brochures Grand Tour in the software mkAlign (Fleury, 2009) so that each string of texts be aligned one to another; hence, both brochures form a bi-text that we can examine contemporaneously: indeed a request on the French part can be mirrored on the Italian part.

3. A Discourse and Translation Study of Adjectives

3.1. Distribution of Adjectives according to their Semantic Categories

Let us first focus on the adjective category. With the CQL request [frpos="ADJ"], the software TXM generates the list of adjectives present in FRTOURISM – the label itpos is used for the Italian part. The TreeTagger language parameter file used for Italian includes in the ADJ category indefinite adjectives (ex. *all*, *each*) that a grammar such as Riegel *et al*. (2009) categorizes as determiners and, in the case of the Italian lemmatization, numerals (cardinals – ex. *due*; ordinals – ex. *primo*). Besides, it does not distinguish between what Biber *et al*. (1999) recognizes as descriptors (denoting the noun's properties or states: color, size, weight, time (chronology, age), emotion and judgements) and classifiers which "delimits or restricts a noun's referent, by placing it in a category in relation to other referents" (*Ibid*.: 508). Therefore, the list must be cleaned before the analysis. For FRTOURISM, out of the 493 items, 288 are hapaxes. We classified the 205 left within Biber's semantic categories and subcategories. We did the same with the ITTOURISM adjectives, where we found 614 items among which 365 hapaxes. Results show that, in terms of items, descriptors and classifiers are equally present in our corpus (FRTOURISM: 103/102; ITTOURISM: 122/127); however, in the French part, descriptive adjectives weigh more in terms of occurrences (531 against 454 classifiers) while the distribution is roughly the same in the Italian part (639 descriptors vs 632 classifiers). The tables below show the five most frequent adjectives for each subcategory, except for colors (only four throughout the corpus).

Table 1. Most frequent adjectives (lemmas) in FRTOURISM (items/occ.)

Descriptors (103/531)					Classifiers (102/454)		
Color (4/19)	Evaluative, Emotive (46/219)	State (16/51)	Size, Quantity, Extent (27/162)	Time (10/80)	Affiliative (18/86)	Relational, Classificational, Restrictive (62/259)	Topical (22/109)
vert	*beau*	*frais*	*grand*	*ancien*	*suisse*	*autre*	*électrique*
rouge	*bon*	*chaud*	*haut*	*nouveau*	*alpin*	*mondial*	*naturel*
blanc	*célèbre*	*doux*	*petit*	*vieux*	*valaisan*	*local*	*industriel*
noir	*idyllique*	*raide*	*nombreux*	*dernier*	*zurichois*	*culturel*	*viticole*
	grandiose	*seul*	*plein*	*actuel*	*bernois*	*central*	*artistique*

Table 2. Most frequent adjectives (lemmas) in ITTOURISM (items/occ.)

Descriptors (122/639)					Classifiers (127/632)		
Color (4/9)	Evaluative, Emotive (59/261)	State (18/66)	Size, Quantity, Extent (34/227)	Time (7/76)	Affiliative (14/108)	Relational, Classifica-tional, Restrictive (95/434)	Topical (18/90)
verde	*bello*	*segreto*	*grande*	*nuovo*	*svizzero*	*culturale*	*naturale*
giallo	*spettacolare*	*aperto*	*alto*	*antico*	*alpino*	*mondiale*	*pubblico*
grigio	*vero*	*fresco*	*piccolo*	*presente*	*tedesco*	*possibile*	*elettrico*
nero	*splendido*	*libero*	*numeroso*	*vecchio*	*bernese*	*urbano*	*industriale*
	buono	*ripido*	*lungo*	*passato*	*francese*	*locale*	*architettonico*

In both languages, relational, classificational and restrictive classifiers on the one hand, evaluative and emotive descriptors on the other are the most frequent subcategories. Such findings are quite coherent with the situational aims brochures must accomplish: as informational texts, they need to describe the reality tourists should discover; as persuasive texts, they also rely quite heavily on the expressions of the emotions this reality provokes. Concerning the evaluative and emotive adjectives, quite numerous are those expressing a positive evaluation felt as above-the-average, even reaching an utmost level. FRTOURISM includes 18 items (68 occ.) while ITTOURISM shows 45 items (125 occ.). The list, in either language, shows how important it is to appeal to senses – touch (*palpitant*), sight (*spectaculaire*; *imponente*) and especially taste (*exquis, savoureux; saporito, gustoso*) – and to judgements of beauty: *sublime, superbe; maestoso, magnifico*. Also, words referring to magic and fantasy are quite numerous: *incroyable, fabuleux; meraviglioso, magico, fantastico, misterioso*. Finally, the list displays a large number of items expressing a unique character that either makes the referent astonishing and/or noticeable (*époustouflant, remarquable, spectaculaire; splendido, esclusivo, eccelente*), or has this referent reach an utmost level: exquis; intenso, perfetto. The exhaustive list informs that Italian tends to resort to and use a larger range of evaluative and emotive adjectives, especially those developing the isotopies of imaginary and uniqueness. Let us now focus on a couple of classifier adjectives whose analysis raises issues in translation: the lemmas *historique/storico. Historique* appears 9 times in FRTOURISM while *storico* 28 in ITTOURISM. A look at their concordances reveals that in Italian, 21 occurrences belong to the collocation *centro storico*. mkAlign, enables us to display, with its mapping functionality, the occurrences of *centro storico* in the *Grand Tour* brochure, and to discover that it matches in French with *vieille ville* (4 occ.) and *coeur historique* (once). In this brochure, the other translation equivalents are the following[1] : s*torici paesini* ⟶ *traditionnels villages, città storiche* ⟶ *sites historiques, edifici storici* ⟶ *architecture historique*. These translations stress the fact that what might be considered as straightforward – the equivalence between *historique* and *storico* – is not so clear-cut. Indeed, depending on its collocations, *historique* calls for various translations: vieux in the case of the collocation vieille ville; *historique* or *traditionnel* in other less fixed expressions. In those cases, the choice between one or the other may be oriented by the writer's need to underline the (rather long) existence of some given points of interest (hence, *historique*) or the memorial features attached to this location (hence, *traditionnels*). A larger corpus

[1] One occurrence of *storico* has no equivalent in French: *otto batelli a vapore storici in stile Belle Époque / huit magnifiques bateaux à vapeur Belle Epoque*.

should help determine the issue. *Historique* and *storico* then appear in utterances focusing on the patrimonial and/or unique characteristics of some given points of interest. Their textual environment points at the referent that the adjective is defining (examples in 1) or indicates something extraordinary that the adjective resumes (examples in 2) – a date in the past, a high altitude, a remarkable quantity (multitude or uniqueness).

(1) Hôtel Weisshorn. **Hôtel historique**. (…) [N°03]
Centre de Fribourg ⟶ Vieille ville. « Funi ». **Transport historique**. Le « Funi » est le dernier funiculaire à contrepoids d'eau de Suisse. [N°03]

(2) Cette **auberge historique** de 1891 trône à 2337 mètres, en surplomb du Val d'Anniviers. [N°03]
Lucerne. (…) ~ 960 **monuments historiques** protégés [N°03]
(…) tous les hébergements du tour dégagent un charme incontestable : qu'il s'agisse (…), d'un **boutique-hôtel historique** ou d'une cabane en rondins avec hot tub extérieur [N°03]
Chi desidera ammirare queste acque in tutta la loro bellezza, può salire a bordo di uno degli otto **battelli a vapore storici** in stile Belle Epoque che da 130 anni attraversano il lago (…) [GT]

Eventually, the concordancers outline the preference of French for generic terms (*architecture, sites, charme*) where Italian uses concrete and/or precise referents – *edifici*, città (3). A larger study will help determine whether this is true for a larger set of tourism discourse or for each language.

(3) (…) non può mancare una passeggiata nel centro storico per conoscere i pittoreschi vicoli e gli edifici storici. [GT] (…)
(…) une promenade dans la vieille ville vous invite à découvrir des ruelles pittoresques et l'architecture historique. [GT]

3.2. Phraseological Segments of Quantity or Intensity

Our second entry is phraseology. Let us focus on two segments. The first one accounts for precisely quantified nominal groups. In FRTOURISM, the request [frpos="NUM"] [frpos="NOM"] [frpos="PRP.*"]? [frpos="DET.*"]?[frpos="NOM|ADJ|NAM"] lists 137 utterances (4). These examples show that the segments accomplish the brochures' pragmatic functions: they give (striking, funny or unusual) historical, geographical or cultural information to get the readers' attention and to turn them into future visitors; they also present practical recommendations to help tourists prepare their visits.

(4) Chacune des **1300 bouteilles d'Intuiva Pinot Noir** du domaine Lipp est unique. [GT]
5 montagnes locales le long du Grand Tour [GT]
Avec (…) **cinq piscines fluviales**, les Zurichois sont particulièrement gâtés (…) [GT]

La visite audioguidée (…) est l'occasion d'écouter (…) raconter **huit siècles d'histoire**. [GT]

(…) la Staubernbahn est le premier téléphérique au monde (…) [N°03]

(...) admirer la Vierge noire du **XVe siècle de la chapelle** des Grâces. [N°03]

In the Italian Grand Tour, these French expressions of quantity find strict equivalents:

(5) Ognuna delle **1300 bottiglie di Pinot Noir Intuiva** della tenuta Lipp è un prezzo unico.

5 montagne di case lungo il Grand Tour

Nel territorio communale si contano sei stabilimenti lungo (…) e **cinque lungo il fiume**.

(…) attraverso l'audioguida alcuni membri (…) raccontano storie appartenenti a **otto secoli**.

The CQL request needs to be refined for future research to encompass cases of ellipsis for example. However, it already puts forward the importance of this phraseological segment at the corpus level, both quantitatively (number of occurrences) and discursively. Another segment that drew our attention involves the use of adverbs: the request [frpos="ADV"] []? [frpos="ADJ"] within 3 lists 124 utterances (after suppression of irrelevant adverbs such as durative ones) illustrated in (6). The most frequent realization includes the superlative form, *le/la/les plus* + ADJ (55 occ.), highlighting the points for which Switzerland breaks records at European or world level. The other utterances correspond to expressions where a quality is valued in terms of quantity (*plus, davantage*) or intensity (*absolument, vraiment*); the adverbs belong to the category of degree (Biber *et al*. 1999), and in all cases amplify or intensify the extent to which the characteristics hold.

(6) (…) le hêtre et le mélèze prennent des couleurs **absolument sublimes**. [N°03]

Le miel (…) a un goût **plus complexe** et présente **davantage de nuances gustatives** (…) [GT]

Les offres de nos partenaires rendront cependant ce voyage **encore plus agréable**. [GT]

Et, du toit praticable du conteneur, la vue est **vraiment panoramique** ! [GT]

(…) tout en faisant ses achats (…), à des prix **magnifiquement avantageux**. [GT]

Découvrez par vous-même d'autres coins et recoins de ce pays **si changeant**. [N°03]

(…) Gandria, le **très romantique** village de pêcheur sur la rive du lac. [N°03]

Presque trop beau pour être vrai ! Coteaux de vignes idylliques (…) [GT]

The Italian *Grand Tour* simplifies the expressions that quantifies or amplifies a judgement: indeed, the repetition of *più* is not avoided in (a) and the intensity is

not translated in (b) to (e). In the case of (e), we notice that, contrarily to the Italian part or our corpus, the French one alters fixed expressions.

(7) (a) Il miele (…) ha un sapore più articolato e offre **più sfumature**
(b) Ovviamente, con le offerte dei nostri partner il viaggio diventa **più gradevole**.
(c) La terrazza realizzata sul tetto del container offre una **vista panoramica a 360°**
(d) (…) acquistare i prodotti di oltre 160 marchi internazionali a **prezzi imbattibili**.
(e) **Troppo bello** per essere vero

4. Conclusion

Our work lays the foundations for a larger study of touristic brochures from a multilingual perspective. It first outlines the fact that, to (re)present the territory the brochures are selling (Switzerland nature, gastronomy and points of interest), our two Italian brochures rely on a large list of euphoric adjectives and develop two main isotopies: fantasy and magic on the one hand, uniqueness and exclusiveness on the other. Our two French brochures appear sober in tone, with a tendency for generic or abstract substantives. A look at phraseological segments leads to an opposite conclusion with respect to sequences where an adverb is followed by an adjective: the Italian brochures tend to simplify or even remove the evaluation that the adverb carries. Therefore, at this stage of the research, we conclude to the fact that, in brochures, the speaker is unidentified in both languages at discourse level but present at the enunciative level, through adjectives in Italian and adverbs in French. From a methodological perspective, the bi-text offers the possibility to describe what happens to texts in the process of translation and which strategies translators adopt: we have seen cases of simplification and normalization in the Italian equivalents of the amplifying phraseological segments. The comparable texts give the opportunity to contrast the characteristics of each monolingual corpus part, free from potential translation effects, as we did with the variety of evaluative and emotive adjectives in Italian. The brochures at our disposal will not only help us test the present conclusions, but also extend them to English and carry a statistical analysis to account for the adjectives' and adverbs' distribution according to the texts' topics and structure.

References

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. (1999). Longman Grammar of Spoken and Written English. Longman.

Carron H. (2015). Caractéristiques textuelles du discours touristique : une étude comparative des substantifs valorisants dans des brochures touristiques wallonnes et flamandes. Université de Gent.

Charaudeau P. (2009). Le discours de manipulation entre persuasion et influence sociale. *Conference Proc*., Lyon. URL: http://www.patrick-charaudeau.com/IMG/pdf/Discours_Manipulation-_Texte_Lyon_.pdf.
Fleury S. (2012). mkAlign. Manuel d'utilisation. Version 2.0.
Heiden S., Magué J.-Ph. and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, pp.1021-1032.
Maci S. M. (2020). English Tourism Discourse. Insights into professional, promotional and digital language of tourism. Ulrico Hoepli Editore.
Nigro M. G. (2006). The Language of Tourism as LSP? A Corpus-Based Study of the Discourse of Guidebooks. In Picht H. editors, *Modern Approaches to Terminological Theories and Applications*. Peter Lang.
Riegel M, Pellat J.-Ch. and Rioul R. (2009). Grammaire méthodique du français. Presse Universitaries de France.

Hybride CNN-SVM : vers une meilleure classification des données manuscrites

Abdallah Benkadja, Ismaïl Biskri, Nadia Ghazzali

Université du Québec à Trois-Rivières – abdellah.benkadja@uqtr.ca
ismail.biskri@uqtr.ca nadia.ghazzali@uqtr.ca

Abstract
We present in this paper a hybrid model combining the CNN convolutional neural network with the SVM classifier. Experiments carried out on handwritten digits and letters from the EMNIST dataset show higher recognition rates than those obtained with CNN and SVM alone. We found by varying some parameters of CNN and SVM that it is possible to further improve the recognition rates of the hybrid CNN-SVM approach.
Keywords: Hybrid CNN-SVM, Classification, EMNIST

Résumé
Nous présentons dans ce papier un modèle hybride combinant le réseau de neurones à convolutions CNN avec le classifieur SVM. Des expérimentations menées sur des chiffres et des lettres manuscrits issus du dataset EMNIST permettent d'observer des taux de reconnaissance supérieurs à ceux obtenus avec CNN et SVM seuls. Nous avons constaté en faisant varier certains paramètres de CNN et de SVM qu'il est possible d'améliorer davantage les taux de reconnaissance de l'approche hybride CNN-SVM.

Mots clés: Méthode hybride CNN-SVM, Classification, EMNIST

1. Introduction

La reconnaissance automatique du manuscrit revêt un intérêt majeur dans le domaine du traitement de l'image et de la reconnaissance des formes. Un plus grand taux de reconnaissance est toujours demandé par la recherche en apprentissage automatique et en intelligence artificielle et même par des applications concrètes en industrie.

De nombreuses études consacrent des approches d'apprentissage automatique dans la reconnaissance des caractères manuscrits et la reconnaissance des formes (LeCun *et al*., 1998; Cortes, Vapnik, 1995; Baldominos *et al*., 2019 ; Liu *et al*., 2003; Keysers, 2007). Parmi les approches les plus utilisées, nous avons les réseaux de neurones à convolution (Convolutional Neural Network) CNN (LeCun *et al*., 1998) et les machines à vecteurs de support (Support Vector Machine) SVM (Cortes, Vapnik, 1995). CNN est une classe de réseaux de

neurones artificiels avancés, largement utilisés dans les problèmes de vision par ordinateur et de traitement de l'image. SVM est une technique d'apprentissage supervisé pour des fins de classification, largement utilisée dans la recherche d'information, la vision par ordinateur, la biologie, etc.

Malgré leur efficacité démontrée dans le domaine de la reconnaissance de forme, plus particulièrement, la reconnaissance des chiffres et des lettres manuscrits, on constate, plus récemment, l'émergence d'une approche hybride combinant CNN et SVM dans laquelle la dernière couche de CNN est remplacée par le classifieur SVM, avec les travaux de (Khairandish *et al*., 2021) sur la classification d'images IRM de tumeurs cérébrales et de classification du manuscrit sur la classification de caractères manuscrits (Ali, Mallaiah, 2021 ; Ahlawat, Choudhary, 2020 ; Niu, Suen, 2012). Dans la reconnaissance des chiffres manuscrits provenant de MNIST (http://yann.lecun.com/exdb/mnist/), Ahlawat et Choudhary (2020), et Niu et Suen (2012) rapportent pour cette approche hybride des taux de reconnaissance respectifs de 98,88 et 99,81, une augmentation respectivement de 1,03 % (98,88 % - 97,85 %) par rapport à SVM seul et de 0,4 % (99,81 % - 99,41 %) par rapport à CNN seul.

Bien que ces travaux tendent à démontrer que Hybride CNN-SVM améliore le taux de reconnaissance des chiffres manuscrits de MNIST, l'avantage de son utilisation, comparativement à CNN et SVM seuls, n'est pas clair quand appliquée à des datasets plus hétérogènes. En effet, la majorité des travaux publiés sur Hybride CNN-SVM donnent des résultats d'expérimentation, seulement, sur les chiffres manuscrits issus de (MNIST).

Notre travail de recherche porte sur une étude comparative des résultats de reconnaissance des caractères manuscrits, à partir de datasets plus variés, obtenus avec CNN, SVM et Hybride CNN-SVM. Un large éventail de paramètres tant pour CNN que pour SVM est considéré.

Les jeux de données utilisés pour l'exploration empirique et l'étude comparative sont ceux de Extended MNIST (EMNIST) (https://www.nist.gov/itl/products-and-services/emnist-dataset) (Cohen *et al*., 2017). Ceux-ci comportent plusieurs ensembles de données qui varient en nombre de classes et en taille d'échantillons d'entrainement et de test. Nous nous limitons, pour les besoins de cette recherche, aux trois jeux de données suivants : EMNIST Letters, EMNIST Balanced et EMNIST MNIST (voir tableau 1). Ces derniers contiennent des données qui ont le même format. Ce sont des images de 28x28 pixels sur une échelle de gris allant de 0 à 255. EMNIST MNIST est un dataset de chiffres, à 10 classes. EMNIST Letters est représenté par 26 classes de caractères latins. On n'y distingue pas les majuscules des minuscules. EMNIST Balanced comprend 47 classes de chiffres et de lettres (pour certaines lettres, on distingue les majuscules des minuscules).

Tableau 1. Datasets utilisés

Dataset	EMNIST MNIST	EMNIST Letters	EMNIST Balanced
Entrainement	60 000	88 800	112 800
Test	10 000	14 800	18 800
Total	70 000	103 600	131 600
Nombre de classes	10	26	47

2. Hybride CNN-SVM

Comme notre travail porte sur Hybride CNN-SVM, nous présentons brièvement SVM et CNN ainsi que leurs différents paramètres en lien avec les expérimentations de notre étude.

SVM vise à obtenir les paramètres de l'hyper-plan ou un ensemble d'hyper-plans dans un espace à dimensions relatives au jeu de données utilisé (Cortes et Vapnik 1995). Dans le travail en cours, le noyau gaussien (RBF) est utilisé sous différentes variations de **γ** et **C**.

γ est un hyperparamètre que nous devons définir avant d'entraîner le modèle. Il détermine la forme de l'hyperplan qui est la frontière de décision. Une valeur élevée de **γ** signifie plus de courbure. Une valeur basse de **γ** signifie moins de courbure. **γ** est obtenu au moyen de la formule **γ=(n*var(X))**, **var(X)** étant la variance du dataset d'entrainement et n le nombre de variables aléatoires représentant les valeurs des pixels des images en entrée. La fonction RBF du noyau gaussien utilisée dans notre travail est définie par l'équation suivante : $\mathbf{exp(-\gamma\|x-x'\|^2)}$. Le paramètre **C** indique la marge de tolérance à l'erreur pour les données d'entrainement. Une valeur élevée de **C** est susceptible de mener à un surapprentissage du modèle qui peut biaiser les résultats de test. Inversement, une très petite valeur de C amènera l'optimiseur à rechercher un hyperplan de séparation à plus grande marge pour les datasets d'entrainement. Enfin la fonction de décision One-Vs-Rest (**OVR**) permet de diviser l'ensemble des données multi-classes en plusieurs problèmes de classification binaire.

Dans le cas de CNN (voir le tableau 2), nous utilisons trois architectures différentes, étiquetées *CNN-1*, *CNN-2* et *CNN-3*. *CNN-1* correspond à la même structure CNN simplifiée à 5 couches qui a été présentée dans (Pan *et al*., 2009). *CNN-2* utilise la même architecture LeNet-5 présentée dans (LeCun *et al*. 1998), à la seule différence que le nombre d'unités de sortie correspond au nombre de classes du dataset d'entrée. *CNN-3* se base sur la même structure que *LeNet-5* avec quelques paramètres modifiés.

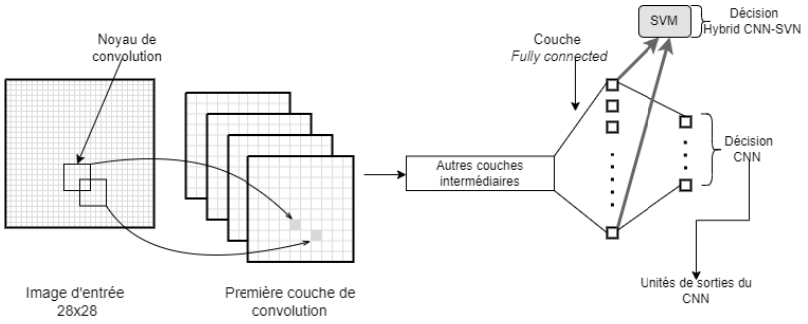
Tableau 2. Architectures CNN utilisées

Couche	CNN-1	CNN-2	CNN-3
1	INPUT	INPUT	INPUT
2	CONV2D filters : 50 kernel : 5 x 5 stride : 2	CONV2D filters : 6 kernel : 5 x 5 stride : 1	CONV2D filters : 32 kernel : 5 x 5 stride : 1
3	CONV2D filters : 50 kernel : 5 x 5 stride : 2	AveragePool	MaxPool stride : 2
4	FC : 100	CONV2D Filters : 16 Kernel : 5 x 5 Stride : 1	CONV2D filters : 48 kernel : 5 x 5 stride : 1
5	FC : OUTPUT	AveragePool	MaxPool Stride : 2
6		FC : 120	FC : 256
7		FC : 84	FC : 84
8 PUT		FC : OUTPUT	FC : OUT-

Nous présentons dans le tableau 2 les différents paramètres de CNN : (i) le *Kernel* qui est une matrice correspondant au noyau de convolution, (ii) le *Stride* qui fait référence au pas que le noyau de convolution considère dans son parcours de la matrice d'entrée, (iii) le *Pooling* dont l'objectif principal est de réduire le nombre d'éléments des matrices de la couche d'entrée, (iv) le *AveragePooling* qui renvoie la moyenne de toutes les valeurs de la partie traitée par le noyau et l'insère dans la position correspondante de la matrice de sortie, (v) le *MaxPooling* qui renvoie la valeur maximale de la partie de la matrice traitée par le noyau et (vi) la couche *Fully connected* (FC) qui sert à convertir la matrice d'entrée en vecteurs entièrement connectés comme un réseau de neurones.

Tel que présenté dans la figure 1, les images d'entrée, normalisées et centrées, représentent la couche d'entrée de notre modèle hybride CNN-SVM. La dernière couche de CNN est remplacée par le classifieur SVM. Ce dernier prend les sorties de la couche *Fully connected* qui précède les sorties de décision de CNN comme un nouveau vecteur de caractéristiques pour l'entrainement par SVM. Les décisions obtenues sont dès lors le résultat de la méthode hybride.

Figure 1. Hybride CNN-SVM



L'objectif du travail étant d'apporter le plus d'éléments de comparaison des résultats de Hybride CNN-SVM à ceux obtenus avec CNN et SVM utilisé chacun individuellement, un large éventail de paramètres tant pour CNN que pour SVM seront expérimentés en leur assignant différentes valeurs. En outre, notre expérimentation empirique s'effectuera sur des jeux de données variés en nombre de classes et en tailles d'échantillons.

3. Expérimentation

Pour les besoins de l'expérimentation, nous avons implémenté le modèle hybride en utilisant les librairies de Keras (Chollet, 2015) et LibSVM (Chang *et al*., 2011). Dans le tableau 3, nous présentons une synthèse des résultats de nos expérimentations ainsi que des résultats trouvés dans la littérature. La première étape de nos expérimentations consiste à effectuer la classification des images manuscrites de nos trois datasets (Tableau 1) au moyen de SVM en faisant varier le paramètre C, et au moyen des trois architectures CNN. Ensuite, nous utilisons le modèle Hybride CNN-SVM en faisant varier, également, le paramètre C et en considérant les trois architectures CNN.

Nous constatons que pour les chiffres manuscrits (MNIST) les taux de reconnaissance sont très élevés (**98,34-99,47%**). Les résultats montrent que CNN-3 a donné le meilleur taux de reconnaissance parmi les 3 autres architectures de CNN. Cette même architecture CNN-3, quand combinée à SVM dans Hybride CNN-SVM permet un meilleur taux de reconnaissance supérieur de 0.71% par rapport à SVM et de **0,10%** par rapport à CNN-3. C'est un constat semblable dans le cas l'Hybride CNN-SVM utilisant CNN-1 et CNN-2. Néanmoins, cette amélioration du taux de reconnaissance reste minime puisque inférieure à 1 %. Dans le cas des lettres (EMNIST Letters), nous observons une augmentation plus significative du taux de reconnaissance de l'approche Hybride CNN-SVM. Cette augmentation est observée peu importe l'architecture CNN utilisée. Hybride-CNN-1-SVM augmente le taux de reconnaissance par rapport à CNN-1 de **1,88%**. Hybride-CNN-2-SVM augmente le taux de reconnaissance par rapport à CNN-2 de **1,28%**. Hybrid-CNN-3-SVM augmente le taux de reconnaissance par rapport à CNN-3 de **1,31%**. L'amélioration du taux de reconnaissance d'Hybride CNN-3-SVM est encore plus significative comparativement à celui de SVM. L'augmentation est supérieure à 2%. Des trois architectures hybrides, Hybride-CNN-3-SVM est celle qui permet d'obtenir le meilleur taux de reconnaissance. Dans le cas de EMNIST Balanced qui contient et des chiffres et des lettres, le nombre de classes est plus important. Cela se traduit empiriquement par un taux de reconnaissance moins élevé. Néanmoins, nous remarquons qu'avec les trois combinaisons hybrides, une augmentation du taux de reconnaissance comparativement à CNN et à SVM. Le meilleur taux de reconnaissance est obtenu avec Hybride-CNN-3-SVM avec 1,57% de plus par

rapport à CNN-3 et 2,6% par rapport à SVM.
Par ailleurs, nous observons que pour l'ensemble de nos datasets, CNN-3 permet d'obtenir des taux de reconnaissance supérieurs à ceux de CNN-1 et CNN-2. Hybride CNN-3-SVM donne, également, les meilleurs taux de reconnaissance comparativement aux variantes Hybrides CNN-1-SVM et CNN-2-SVM. Cela démontre que les performances du modèle hybride varient en fonction des paramètres de ses composantes que sont CNN et SVM. Globalement, l'approche l'Hybride CNN-SVM est empiriquement avantageuse puisqu'elle tend, globalement, selon nos expérimentations, à donner des résultats meilleurs que CNN et SVM.
Biswas et Islam (2021) ont obtenu un taux de reconnaissance des chiffres au moyen de CNN égal à 99,53%. Cavalin et Oliviera (2019) ont obtenu un taux de reconnaissance des lettres au moyen de CNN égal à 93,63%. Nos meilleurs résultats avec Hybride CNN-SVM leur sont légèrement inférieurs. Néanmoins, quand appliqué à un jeu de données plus large et plus hétérogène, notre approche hybride donne un meilleur résultat que celui de Cavalin et Oliveira avec 88,22% de taux de reconnaissance contre 87,18%.

Tableau 3. Résultats des expérimentations (en %)

Technique	**Paramètre C de SVM**	**MNIST (Chiffres)**	**Letters (Lettres)**	**Balanced**
SVM	1	98,34	88,98	84,28
	3	98,73	90,10	85,52
	5	**98,76**	**90,29**	**85,62**
CNN-1		99,10	90,50	85,38
CNN-2		98,96	90,97	86,38
CNN-3		**99,37**	**91,53**	**86,65**
Hybride CNN-1-SVM	1	99,30	92,38	87,89
	3	99,27	92,32	87,71
	5	99,20	92,23	87,49
Hybride CNN-2-SVM	1	99,16	92,20	87,87
	3	99,23	92,22	88,09
	5	99,20	92,25	88,08
Hybride CNN-3-SVM	1	99,43	92,79	**88,22**
	3	**99,47**	**92,84**	88,03
	5	**99,47**	92,66	87,91
CNN (Biswas, Islam, 2021)		99,53		
SVM (Ahlawat, Choudhary, 2020)		97,85		
CNN (Cavalin, Oliveira, 2019)		99,46	93,63	87,18
Hybride CNN-SVM (Ahlawat, Choudhary, 2020)		98,88		
Hybride CNN-SVM (Niu, Suen, 2012)		99,81		

4. Conclusion et perspectives

Le modèle Hybride CNN-SVM proposé dans ce travail est abordé selon deux perspectives : Sa capacité à reconnaitre les caractères manuscrits et celle à maintenir sa performance en termes de taux de reconnaissance par rapport à CNN et à SVM. Notre approche utilise CNN dans l'extraction des descripteurs des caractères manuscrits avant d'utiliser SVM comme classifieur final. Nos expérimentations ont été menées sur des données variées composées de chiffres et de lettres issues de EMNIST.

Parmi toutes les combinaisons expérimentées, nos résultats empiriques montrent que l'approche Hybride CNN-3-SVM donne les meilleures classifications pour tous nos datasets. Une expérimentation et une validation avec des ensembles de données plus complexes et plus variées pourrait encore davantage consolider les résultats de sa performance.

Bien entendu, nous pensons que l'expérimentation avec la variation des paramètres de SVM et CNN reste au centre de l'amélioration des performances du modèle hybride. Nous en avons exploré quelques-uns dans ce travail. D'autres seront considérés dans nos travaux futur. Parmi ceux-là, il y a les paramètres des couches intermédiaires de CNN, et ceux du noyau SVM, etc.

Bibliographie

Ahlawat S. et Choudhary A. (2020). "Hybrid CNN-SVM Classifier for Handwritten Digit Recognition." In *Procedia Computer Science*, Volume 167. Elsevier. 2554-2560.

Ali A.A.A. et Mallaiah S. (2021). Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout. *Journal of King Saud University-Computer and Information Sciences*.

Baldominos A., Saez Y. et Isasi P. (2019). "A Survey of Handwritten Character Recognition with MNIST and EMNIST." *Journal of Applied Sciences* 9(15). MDPI.

Biswas A. et Islam M. S. (2021). "An Efficient CNN Model for Automated Digital Handwritten Digit Classification." *Journal of Information Systems Engineering and Business Intelligence*, 7(1), 42-55.

Cavalin P. et Oliveira L. (2019). "Confusion Matrix-Based Building of Hierarchical Classification". In proceedings of the 2018 Iberoamerican Congress on Pattern Recognition. LNCS. Springer. 271–78.

Chang C. C. et Lin C. J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 1-27.

Chollet F. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Cohen G., Afshar S., Tapson J. et Van Schaik A. (2017). "EMNIST: Extending MNIST to Handwritten Letters." In *Proceedings of the International Joint Conference on Neural Networks*. pp. 2921-2926

Cortes C. et Vapnik V. (1995). "Support-Vector Networks." *Machine Learning* 20(3): 273–97. Springer.

Khairandish M.O., Sharma M., Jain V., Chatterjee J. M. et Jhanjhi N. Z. (2021). "A Hybrid CNN-SVM Threshold Segmentation Approach for Tumor Detection and

Classification of MRI Brain Images." *IRBM*. Elsevier.
Keysers D. (2007). “Comparison and Combination of State-of-the-Art Techniques for Handwritten Character Recognition: Topping the MNIST Benchmark.” arXiv preprint *arXiv:0710.2231*
LeCun Y., Cortes C. et Burges C. J. C. (1998). “MNIST Handwritten Digit Database” http://yann.lecun.com/exdb/mnist/.
Liu C. L., Nakashima K., Sako H. et Fujisawa H. (2003). “Handwritten Digit Recognition: Benchmarking of State-of-the-Art Techniques.” *Pattern Recognition* 36(10). Elsevier. 2271–2285.
Niu X. X. et Suen C. Y. (2012). “A Novel Hybrid CNN-SVM Classifier for Recognizing Handwritten Digits.” *Pattern Recognition* 45(4). Elsevier. 1318–1325.
Pan W. M., Bui T. D. et Suen C.Y. (2009). “Isolated Handwritten Farsi Numerals Recognition Using Sparse and Over-Complete Representations.” In *Proceedings of the International Conference on Document Analysis and Recognition*, ICDAR, IEEE Computer Society, 486–90.

Emotional Text-Mining in pediatria: Esplorazione del self-care nei bambini e nei giovani in condizioni croniche complesse

Valentina Biagioli[1], Francesca Greco[2], Giuseppina Spitaletta[1], Annachiara Liburdi[1], Rachele Mascolo[1], Orsola Gawronski[1], Riccardo Ricci[1], Emanuela Tiozzo[1], Ercole Vellone[3], Teresa Grimaldi Capitello[1], Michele Salata[1], Massimiliano Raponi[1], Immacolata Dall'Oglio[1]

[1]IRCCS Ospedale Pediatrico Bambino Gesù – valentina.biagioli@opbg.net pinaspitaletta@virgilio.it annachiaraliburdi@gmail.com rachele.mascolo@opbg.net orsola.gawronski@opbg.net riccardo.ricci@opbg.net emanuela.tiozzo@opbg.net teresa.grimaldi@opbg.net michele.salata@opbg.net massimiliano.raponi@opbg.net immacolata.dalloglio@opbg.net
[2]Università di Roma "La Sapienza" – francesca.greco@uniroma1.it
[3]Università di Roma "Tor Vergata" – ercole.vellone@uniroma2.it

Abstract
This study describes the application of text mining techniques in the complex chronic disease field. We enrolled 104 participants, among patients from 6 months to 24 years old, parents, siblings, healthcare professionals, and operators of life contexts. Individual interviews and on-line focus groups were conducted and the contents were transcribed verbatim.The texts were collected in a large size corpus (n. token = 176,824) with a rich vocabulary and it was analyzed using the Emotional Text Mining (ETM) method. The ETM, using the associative links between the words, detected the general topics and the cultural - symbolic categories influencing the self-care behaviors of the respondents. The results show both the critical issues and the useful resources to identify, plan, and provide personalized services to the population in different contexts.

Keywords: Self-care, Chronic disease, Pediatric, Young people, Health Care Professionals network, Emotional text-mining

Riassunto
Questo lavoro descrive l'utilizzo delle tecniche di text mining nell'ambito delle malattie croniche complesse. Hanno partecipato allo studio 104 soggetti tra pazienti dai 6 mesi e 24 anni, i loro genitori o fratelli e sorelle, i professionisti sanitari e operatori connessi

ai loro contesti di vita. Sono state condotte interviste individuali e focus group on-line, trascritti verbatim. I testi, sono stati raccolti in un corpus di grandi dimensioni (n. token = 176.824) che è stato analizzato col metodo dell'Emotional Text Mining. Sulla base dell'associazione delle parole sono state individuate le tematiche generali e le categorie simbolico-culturali che orientano i comportamenti di self-care degli intervistati. I risultati hanno consentito di evidenziare le criticità e le risorse utili a identificare, pianificare e fornire servizi che siano il più individualizzati possibile nei diversi contesti di vita del paziente.

Parole chiave: Self-care, Patologie corniche, Pediatria, Giovani adulti, Rete di cura dei professionisti sanitari, Emotional text-mining

1. Introduzione

In Italia, i bambini in condizioni croniche rappresentano circa il 10% della popolazione (Ministero della Salute, 2010). Secondo Cohen *et al*. (2011), la complessità medica dei bambini e giovani è caratterizzata da: bisogni di servizi sanitari, una o più condizioni cliniche croniche, diagnosticate o sconosciute, gravi limitazioni funzionali ed elevato utilizzo delle risorse sanitarie. Il coinvolgimento attivo dei pazienti pediatrici e delle loro famiglie nella gestione della loro condizione è quindi fondamentale (Dall'Oglio *et al*., 2019; Smith e Coleman, 2009). Il Piano Nazionale della Cronicità, infatti, orienta allo sviluppo di modelli che garantiscano un approccio integrato per soddisfare i bisogni di salute, soprattutto a domicilio, considerando la peculiarità dell'età evolutiva (Ministero della Salute, 2016).

Il self-care è definito come la capacità degli individui, delle famiglie e delle comunità di promuovere la salute, prevenire le malattie, mantenere la salute e affrontare la malattia e la disabilità con o senza il supporto di un operatore sanitario (WHO, 2021). Riegel *et al*. (2008) hanno definito il self-care come un processo decisionale - basato sull'esperienza (Riegel e Dickson, 2008). In particolare, hanno identificato tre aspetti: self-care maintenance – i comportamenti utilizzati dagli individui con malattie croniche per mantenere la stabilità della malattia da un punto di vista fisico ed emotivo; self-care monitoring - il processo di osservazione di sé stessi per individuare eventuali cambiamenti di segni e sintomi; self-care management - la risposta a segni e sintomi quando si verificano (Riegel e Dickson, 2008; Riegel e Moser, 2018). Pertanto, un approccio di promozione del self-care, iniziato già durante l'infanzia con il sostegno dei genitori e modulato durante la crescita, è fondamentale per garantire la salute nel corso delle diverse fasi di sviluppo. L'obiettivo di questo studio è stato quello di esplorare i comportamenti di self-care nei bambini e giovani in condizioni croniche complesse e il contributo al self-care fornito dalla famiglia e dalla rete del contesto di vita (genitori, fratelli e sorelle, professionisti sanitari

e insegnanti). Inoltre, lo studio ha voluto esplorare i principali fattori che influenzano il processo di self-care durante l'età evolutiva.

2. Materiali e metodi

2.1. Partecipanti

Nel presente studio qualitativo, sono stati arruolati i bambini e giovani fino a 24 anni d'età in condizioni croniche complesse, come descritto da Cohen e colleghi (2011), con una diagnosi da almeno sei mesi. Sono stati arruolati anche i genitori e/o i fratelli e sorelle; i professionisti sanitari che quotidianamente si prendono cura di loro sia in ospedale che nel territorio, ed altre persone connesse all'ambiente di vita dei bambini/ragazzi in studio (scuola, sport, etc.). Tutti i pazienti coinvolti erano assistiti dai diversi servizi dell'IRCCS Ospedale Pediatrico Bambino Gesù.

2.2. Raccolta dati

I partecipanti venivano arruolati sia online che in presenza, previa firma del consenso informato. Contestualmente, veniva compilata una scheda per raccogliere le informazioni cliniche e socio-anagrafiche dei pazienti e dell'eventuale genitore o fratello/sorella partecipante. Inoltre, per i professionisti sanitari e gli operatori dei contesti di vita venivano raccolte informazioni a carattere socio-professionale. I focus group e le interviste sono stati condotti in videoconferenza online, tramite una piattaforma facilmente accessibile a tutti, in ragioni delle limitazioni imposte dalla pandemia. I focus group online hanno consentito ai partecipanti di unirsi alla discussione contemporaneamente e di rispondere alle domande, nonché ai commenti reciproci, in tempo reale. Ciascun focus group è stato svolto da un moderatore, che poneva le domande stimolo, con l'aiuto di un osservatore che raccoglieva le note principali. Invece le interviste sono state svolte da un moderatore, mentre l'osservatore non partecipava attivamente. Nel corso dei focus group sono state poste alcune domande aperte pertinenti con gli obiettivi dello studio. Le domande sono state opportunatamente modulate e semplificate al fine di facilitarne la comprensione da parte delle diverse tipologie di partecipanti.

2.3. Analisi dei dati

Le interviste e i focus group sono stati registrati ed il contenuto è stato trascritto verbatim, utilizzando un sistema di trascrizione automatica on-line, coadiuvata da una verifica ed eventuale modifica o integrazione da parte del ricercatore. Il corpus testuale di grandi dimensioni è stato analizzato con il metodo dell'Emotional Text Mining (ETM) (Greco, 2016; Greco e Polli, 2020). Tale metodologia di Natural Language Processing (NPL) si colloca in una prospettiva di matrice socio-costruttivista a orientamento psicodinamico. L'ETM, attraverso

un approccio sia semiotico che semantico al testo, consente di identificare tre livelli della comunicazione: gli assi generali della comunicazione, i temi principali e i sottotemi (Zolnoori *et al*., 2019).
Inizialmente i testi sono stati pretrattati e sono stati calcolati tre indicatori lessicali per valutare la possibilità di processare statisticamente i dati: i token, il type-token ratio e la percentuale di hapax (Giuliano e La Rocca, 2010). Il corpus è stato suddiviso in unità di contesto, lemmatizzato e sono state escluse le stop word e i termini di rango alto e basso di frequenza (Bolasco, 1999) al fine di ridurre la dimensione della matrice (Greco, 2016). Sulla matrice termini-unità di contesto è stata effettuata un'analisi dei cluster con un algoritmo bisecting k-mean, limitato a 10 partizioni, con l'esclusione dei testi che non presentavano almeno una co-occorrenza, ed è stato utilizzato il coseno come misura di similarità (Steinbach *et al*., 2000). Il Calinski-Harabasz, il Davies-Bouldin e il coefficiente di correlazione intraclasse sono stati calcolati per individuare le prime due partizioni ottimali. La prima partizione ottimale ha consentito di individuare le macrotematiche mentre la seconda partizione ottimale ha permesso di individuare le sottotematiche (cluster) all'interno delle macrotematiche. Successivamente, sulla matrice termini-cluster della prima partizione ottimale è stata effettuata una analisi delle corrispondenze (Lebart e Salem, 1994) per identificare gli assi principali della comunicazione (Greco, 2016). Infine, sulla tabella di contingenza cluster-tipo di partecipante sono stati calcolati il χ^2 e i residui standardizzati per valutare le differenze nella rappresentazione della auto-cura (Sharpe, 2015).

3. Risultati

La numerosità campionaria complessiva è di 104 partecipanti. In particolare, sono stati arruolati: 27 pazienti; 33 genitori divisi per fascia d'età; 6 fratelli/sorelle di pazienti che sono stati suddivisi in due gruppi in base all'età; 33 professionisti sanitari (infermieri, pediatri specialisti, dietisti, psicologi, fisioterapisti); 5 tra professionisti od operatori che svolgono la loro attività nei contesti educativi e di vita della popolazione oggetto dello studio.
I 364 termini selezionati hanno consentito di classificare il 99,6% delle unità di contesto e sono state individuate due partizioni ottimali: una prima a cinque cluster (macrotematiche) e una seconda a nove cluster (sottotematiche). Infine, l'analisi delle corrispondenze ha individuato quattro fattori (assi della comunicazione) di cui i primi tre spiegano l'82,3% dell'inerzia.
I partecipanti hanno rappresentato il processo di self-care attraverso quattro assi principali: Cura di sé, Setting esterni, Famiglia e Gestione della malattia. Il primo fattore (inerzia= 38,7%) ha evidenziato il processo di "Self-care" come l'impatto della "condizione cronica" sulla vita dei pazienti e delle famiglie (*polo positivo*) e i "comportamenti" (*polo negativo*) che i pazienti eseguono durante le loro

attività della vita quotidiana.
Il secondo fattore (inerzia=23,7%) era rappresentato dai 'Setting esterni' in cui i pazienti e le famiglie sono impegnati in attività di self-care. In particolare, i partecipanti hanno distinto l''ospedale' (*polo positivo*), come principale punto di riferimento per i pazienti e le famiglie, dalla 'scuola' (*polo negativo*), come il setting sociale dove i pazienti applicano autonomamente comportamenti di self-care come l'assunzione di farmaci.
Il terzo fattore (inerzia=19,8%) ha evidenziato il ruolo essenziale della 'Famiglia' percepita come una risorsa globale di cura nel tempo. Da un lato, i partecipanti si sono concentrati sulla 'gestione della condizione cronica' (*polo positivo*) percepita non solo come le attività che i pazienti e le loro famiglie svolgono a casa, ma anche come la loro risposta sociale ed emotiva alla condizione di cronicità. D'altra parte, la famiglia rappresentava il contesto dove poter svolgere una "vita normale" (*polo negativo*), per quanto possibile.
Infine, il quarto fattore (inerzia=17,7%) si è concentrato sulla 'Gestione' della condizione cronica sia in modo dipendente che autonomo. I partecipanti hanno distinto l'"autogestione" (*polo positivo*), caratterizzata dall'importanza dell'empowerment dei pazienti nell'esperienza di cura durante l'età evolutiva, dal "supporto professionale" (*polo negativo*), fornito da un team multidisciplinare sia nel setting ospedaliero che nella comunità. Il supporto professionale è stato considerato fondamentale per aiutare i pazienti ad affrontare la loro condizione di cronicità.
Sono pertanto emerse cinque macrotematiche e nove sottotematiche: 1) il self-care management suddiviso in 'la gestione dei dispositivi' e 'il consulto professionale'; 2) il passaggio di responsabilità che distingue 'i fattori influenzanti', 'i genitori' e 'la scuola'; 3) il supporto al self-care che si declina in 'vita normale' e 'supporto multidisciplinare'; 4) il self-care nella vita quotidiana e 5) l'aderenza al trattamento.
Queste macrotematiche riflettono il processo di self-care nei bambini e i giovani in condizioni croniche complesse. Questo processo è caratterizzato da comportamenti di self-care come: il self-care maintenance (aderenza al trattamento), il monitoraggio dei segni e sintomi, e la gestione delle complicanze. Il self-care nei suoi tre livelli si evolve lungo un graduale passaggio di responsabilità dai genitori al paziente. Inoltre, il processo di self-care è influenzato dal supporto multidisciplinare dei professionisti e degli operatori del contesto di vita.
Infine, risulta significativa l'associazione tra la macrotematica e il tipo di intervistato (χ^2, df = 16, $p < 0{,}001$). Genitori e pazienti hanno una esperienza simile di self-care: si concentrano principalmente sulla gestione delle complicanze, il self-care maintenance e il monitoraggio dei segni e sintomi nella vita quotidiana e sull'aderenza al trattamento. Sembrerebbe che non riportino o

non diano risalto alle tematiche legate al passaggio di responsabilità dai genitori al paziente ed al supporto multidisciplinare. I fratelli differiscono leggermente dai precedenti, concentrandosi prevalentemente sul self-care maintenance e sul monitoraggio dei segni e sintomi nella vita quotidiana e sull'aderenza al trattamento.
Al contrario, i professionisti sanitari e gli operatori dei contesti di vita si concentrano principalmente sul passaggio di responsabilità dai genitori al paziente, prescindendo, dal self-care maintenance, dal monitoraggio dei segni e sintomi nella vita quotidiana e dalla gestione delle complicanze. Anche i professionisti sanitari non si concentrano sull'aderenza al trattamento, e si differenziano dagli operatori dei contesti di vita in quanto si interessano molto anche al supporto e all'educazione al self-care.

4. Discussione e conclusioni

Le cinque macrotematiche emerse dall'analisi sintetizzano il processo di self-care messo in atto dai bambini e giovani in condizioni croniche complesse durante la loro vita quotidiana. I professionisti sanitari dovrebbero promuovere il self-care tramite l'educazione alle famiglie dei bambini e giovani in condizioni croniche complesse. Inoltre sarebbe utile adattare il trattamento terapeutico allo stile di vita del bambino/ragazzo affinché questo non rappresenti un ulteriore onere, riducendo così anche il rischio di non aderenza alla terapia.
Per ridurre l'onere dei genitori, che rimangono la figura di riferimento per un adeguato self-care, sarebbe opportuno implementare strategie di coping e programmi di supporto psicologico, al fine di diminuire l'impatto che la malattia ha sulle loro vite e migliorare la loro qualità di vita. Inoltre sembra emergere il valore aggiunto di una rete di assistenza domiciliare integrata e di un infermiere case manager nel garantire la continuità assistenziale, la sicurezza e la tranquillità delle famiglie.
In conclusione, per un'assistenza incentrata sul paziente, sarebbe opportuno conoscere a fondo le esperienze dei bambini e delle famiglie che si trovano a dover convivere con una condizione cronica complessa e utilizzare tali conoscenze per pianificare e fornire servizi il più individualizzati possibile.

Bibliografia

Bolasco S. (1999). *Analisi multidimensionale dei dati: Metodi, strategie e criteri d'interpretazione*. Carocci.
Cohen E., Kuo D.Z., Agrawal R., Berry J.G., Bhagat S.K., Simon T.D. e Srivastava R. (2011). Children with medical complexity: An emerging population for clinical e research initiatives. *Pediatrics*, 127(3):529–538.
Cordella B., Greco F. e 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 e S. Fleury, editors, Proceedings JADT 2014, *12es Journées*

internationales d'Analyse Statistque des Données Textuelles, Paris, France, Juin 3-6, 2014, pp. 173-184.

Dall'Oglio I., Mascolo R., Tiozzo E., Portanova A., Fiori M., Gawronski O., Dotta, A., Piga, S., Offidani, C., Alvaro, R., Rocco, G., Latour, J.M. e FCC Italian NICUs Study Group. (2019). The current practice of family-centred care in Italian neonatal intensive care units: A multicentre descriptive study. *Intensive e Critical Care Nursing*, 50:36–43.

Giuliano L. e 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. e 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. e Salem A. (1994). *Statistique Textuelle*. Dunod.

Ministero della Salute (2010). Relazione sullo Stato Sanitario del Paese 2009-2010 [web site]. (https://www.quotidianosanita.it/allegati/allegato9849531.pdf, accessed 27 July 2021).

Ministero della Salute. (2016). Piano nazionale delle cronicità [web site]. (http://www.salute.gov.it/imgs/C_17_pubblicazioni_2584_allegato.pdf, accessed 30 April 2021).

Riegel B. e Dickson VV. (2008). A situation-specific theory of heart failure self-care. *The Journal of Cardiovascular Nursing*, 23(3):190–196.

Riegel B. e Moser D (2018). Self-care: An Update on the State of the Science One Decade Later. *The Journal of Cardiovascular Nursing*, 33(5):404–407.

Sharpe D. (2015). Your Chi square test is statistically significant: Now what? *Practical Assessment, Research e Evaluation*, 20(8):1–10.

Smith L. e Coleman V. (2009). *Child e Family-Centred Healthcare: Concept, Theory e Practice*. Macmillan International Higher Education.

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

WHO (2021). What do we mean by self-care? [web site].
(http://www.who.int/reproductivehealth/self-care-interventions/definitions/en/, accessed 27 July 2021).

Zolnoori M., Balls-Berry J.E., Brockman T.A., Patten C.A., Huang M. e Yao L. (2019). A systematic framework for analyzing patient-generated narrative data: Protocol for a content analysis. *JMIR Research Protocols*, 8(8):13914.

Impact of the expression of feelings and emotions in fake news and their reception: Specific electronic dictionary and automatic recognition with NooJ

Magali Bigey

University of Franche-Comté – magali.bigey@gmail.com

Abstract
The power of the image is to touch the deepest feelings of the public (indignation, compassion, sublimation, fear, hatred, amazement, etc.). Internet users, who are often drowned in a permanent flow of information, are not led to analyze an image. They are simply touched, affected by the content. This appeal to emotion is a favorable ground for the deployment of post-truth. After having made this observation, we are interested in the construction of fake news discourses and their form, suggesting a grammar of the infox. A study of this grammar of infox shows us the importance of a narrative coherence, which largely takes precedence over an argumentative coherence and constructed reasoning. We also found, during a reception study, that feelings and emotions were important levers of permeability to fake news and particularly to infox by image. We will show how to automatically analyze the reception of fake news and infox discourses through images with the NooJ software, using added semantic information and local grammars, and also showing the interest of over-coding dictionaries.

Keywords: Reception analysis, Infox, Fake News, Emotions, Electronic dictionary.

1. Introduction

The notion of opinion, like that of feelings, remains vague when it comes to expression on the Internet. It is necessary to define the limits of each of the parameters that come into play when thinking about analysis, and not to take every text for an expression of opinion, often coupled with feelings. The power of the image is to touch the deepest feelings of the public (indignation, compassion, sublimation, fear, hatred, amazement, etc.). Internet users, often drowned in a permanent flow of information, are not led to analyze, are most of the time touched, affected by the content. This appeal to emotion is a favorable ground for the deployment of post-truth (where the reality of the facts influences public opinion less than the appeal to emotions and personal beliefs). The emotion that prevails, truth has become secondary, fertile ground for the circulation of all kinds of infoxes.

2. Expressing one's opinion

We know that feelings and emotions are important levers in the inscription of oneself in discourses, and in the permeability to fake news and especially to infox by image (Bigey & Simon, 2021).

"An utterance is therefore always the realization of an act of enunciation. This thesis [...] implies that the statement [...] must not only be referred to the state of things, to the events, to the facts it reports, but to the conditions of its enunciation, context and production, recipients of the message, intentions of the sender, which determine its form and meanings" . It is very important to be aware of this fact, otherwise the analysis can easily be cut off from information essential to its proper realization. It is essential to examine the context as closely as possible, especially since, as we have seen almost daily, the discourse on the web provokes a kind of emulation that leads to an overbidding of opinions and criticisms, of the meanings of the situations expressed:

"These meanings result from an activity that makes representation a ″construction″ and an ″expression″ of the subject" .

The expression of the subject is here totally linked to its emotions, to what is called the involvement of the self in the discourse.

Here for what concerns infox and post-truth, the expression of the subject enters in adequacy with the collective image in a posture of the between-self. For a given community, each member feels attacked by the critics, all the more so if this entre-soi is the rule, as we regularly observe on social networks.

The expression of feelings "in public"?

For David Le Breton,

"If socially it is hardly conceivable to give free rein to certain emotions, the appropriate places authorize on the other hand to live oneself without detour, without fear of the judgment of the others, without need to curb one's sensitivity.

We can easily read in the comments of the socio-numerical networks, in the posts examples of self-assertion and claims. For Claudine Haroche, we find in what she calls "the forms of contemporary alienation" the idea of self-visibility:

"a type of visibility which, ignoring the borders of the intimate, the private and the public, tends to instrumentalize and reify the individual by the continuous and exhaustive exhibition of oneself, encouraging and reinforcing the voyeurism, the exhibitionism, the erasure of the private, of the intimate [... Reinforced by contemporary technologies, this alienation would force the individual to present not a part of himself but to give himself up to a nakedness, a continuous unveiling of himself [...] The visibility would be synonymous of legitimacy, utility, guarantee of quality: the frequency, the quantity or even the continuity of visibility would value the individual. Invisibility, on the other hand, would be synonymous with uselessness, insignificance or even non-existence.

All these elements, adequacy of the place, visibility of oneself synonymous of

legitimacy, continuous unveiling of oneself, give to think that the Web is the ideal place to unveil oneself, to expose one's feelings, to be visible and to unveil too much, often, all the time. Seen as an agglomeration of "social spaces welcoming the expression of feelings that could not openly be lived elsewhere", the web is therefore a unique place (non-place?) of study.
In a second step, we will show how to automatically analyze fake news and infox reception discourses with the NooJ software, using added semantic information and local grammars, and also showing the interest of dictionary overcoding. We will work on lexical and concordance analyses, but also on analyses of expressed feelings. We will show, from a reception study with first year students of DUT Information-Communication, the importance of emotions in the reception of infoxes. We will make the link aiming at understanding their (in-)capacity of discernment of infoxes in a digital context, their verbalization of personal emotions and the forms of implication of oneself in the discourse. We will also be interested in the hapaxes present on certain sites dedicated to Fake News.

3. Which methodology?

From a methodological point of view, the study of feelings seems difficult because they are complex, structurally and linguistically speaking.
"Feelings and emotions are not absolute states, substances that can be transposed from one individual and one group to another [...] Each term of the affective lexicon of a society or a social group must be related to the local context of its concrete implementation. [In a context of comparison between cultures, the use of affective terms imposes to always put them in quotation marks to recall the vagueness which surrounds them, or to use the vernacular terms to underline that such a comparison is not obvious and remains a question".
Concerning this idea of the categorization of the emotions, it is obviously very difficult to list them all, so much the manners to say them are diverse and varied. Sometimes, it can be a question of a lack of precision, but also of vocabulary spread in a post or a comment, which will give this effect of empathy or not to the reader without it being possible to systematically and automatically identify what generated it. Even if we could imagine a harmonization of the expression of feelings and emotions, we can imagine differences according to the social categories, the experience, the tastes... of each one: one will like a lot, the other will adore, the other will appreciate infinitely, but to what degree, on what scale? The standardization of the analysis is impossible, as well as the standardization of the meaning, so we based ourselves on the study surrounding the verbs of feelings, with a methodology combining lexicometry and automatic processing.

3.1. For a dictionary of sentimental verbs: Automatic recognition of semantic information

There are many methods to analyze large text corpora and to extract semantically homogeneous information from them. Here, we have chosen to work with specific electronic dictionaries, to which we have added semantic information (for example, a V+sent code to recognize sentiment verbs).

Often, automatic processing software is able to recognize verbs, but they do not usually make any semantic distinction, they are based on the form of a word for its recognition. If it is possible to extract all the verbs from a corpus, it is then more difficult to sort them out, and what we are interested in would be drowned in a flood of useless information for the analysis; we have therefore decided to create a specific electronic dictionary that we will call "Vsent" ("Vsent" for Verbs of feelings, this will also be the code applied to each of the verbs of the dictionary). The software used for this is NooJ; it will allow us to extract everything that has been identified as a feeling verb. Obviously, in studies based on meaning, a return to the text is essential, in order to avoid the pitfall of arbitrary choices, raised by Jean-Marie Viprey (1997)

"In this way, one will inevitably end up integrating black with the theme of death, white with that of virginity, black and white together with the themes of virginity, racism, ethics, cinema, etc.; that is to say, in fact, by projecting onto the sometimes spidery weave of the text the mathematical apparatus of thematic associations.

The objective is that, through the software, one can recognize these verbs in particular and extract them in context in the form of concordances in order to perform an analysis.

3.2. Creation of the dictionary

The interest of such a dictionary is to be able to create a concordancer gathering all the verbs of feelings of a corpus in context, and thus, to analyze the whole as a new corpus. For this dictionary creation, we have chosen to base ourselves on tables 4 and 12 of the verbs of the lexicon-grammar of M. Gross (1975), which list the verbs that are felt by a subject, and the second one the verbs of value judgment (to love, to adore, to admire, to hate...). In these tables, correspondences between the syntactic and semantic criteria of certain verbs have been discovered, and this is the case for most of the verbs that interest us. In the form of tables, this classification systematically lists the syntactic and semantic properties of verbs. M. Gross lists the properties of about 3000 verbs, and each table details one or two syntactic constructions and properties. It is thus from these different properties that Mr. Gross created his classes. It turned out that verbs with identical syntactic properties often had similar or even identical semantic properties.

Table 4
For M. Gross, table 4 can be considered as a table of feeling verbs: "the verbs in this table are semantically homogeneous. The vast majority of them correspond to a feeling "triggered" by N0 and "experienced" by N1". Hence our interest in this particular table.
It is presented in the form of a table. The main properties represented are those of the subject completive as well as those of the associated adjectival form N0 is V-a for N1. This table lists several hundred verbs.
Table 12
It can be considered as a table of value judgement verbs. The verbs that make it up are love, adore, admire, hate, despise, envy, hate... For M. Gross, "the verbs in this table [...] are semantically homogeneous and generally trigger a value judgment". For the verbs in table 12, all subjects are human and all direct complements that are nouns are also human nouns.
Verbs enter into operator constructions of the type:
N0 has Det V-n for N1
There are five different subject constructions and there are also 25 different complement constructions, some of them accepting only a few verbs on the whole table. This is the case for the construction N1 is Det V-n which accepts only abominate, bless, abhor, require.
For a dictionary of feeling verbs
The interest of such a dictionary allows us to work only on the verbs that interest us. First, we have listed all the verbs of feelings in tables 4 and 12, and we have entered them in order to obtain a list. As for the coding of the dictionary, we chose simple codes:
V: verb
sent: of feeling
+4: from table 4 of the lexicon - grammar.
+12: from table 12 of the lexicon - grammar.
The advantage of these last two codes is twofold. Firstly, they allow to recognize verbs of feelings among other verbs, and secondly, they allow to recognize verbs from table 4 among verbs from other tables, with a view to creating a complete electronic dictionary of semantically homogeneous tables of the lexicon - grammar.
The first dictionary thus obtained includes 538 verbs,
Once saved as a new file with its codes in a plain text format in NooJ, the dictionary is usable. It has been compiled so that the inflection codes are automatically inserted during the automatic analysis of the corpus.
The sentiment verbs thus identified are easily extracted in the form of a concordancer, with a "V+sent" query.
Example of infox reception analysis:

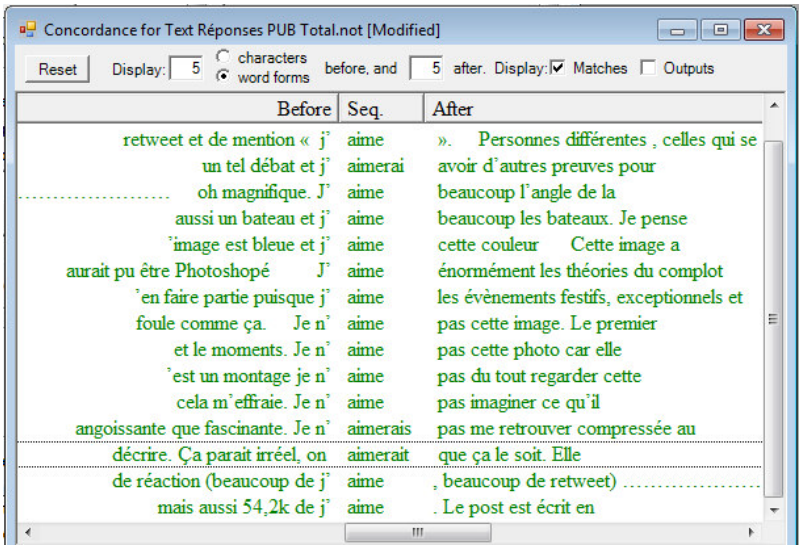


Figure 1. Concordance for Vsent Verbs

Coupled with markers of self-involvement in discourse:

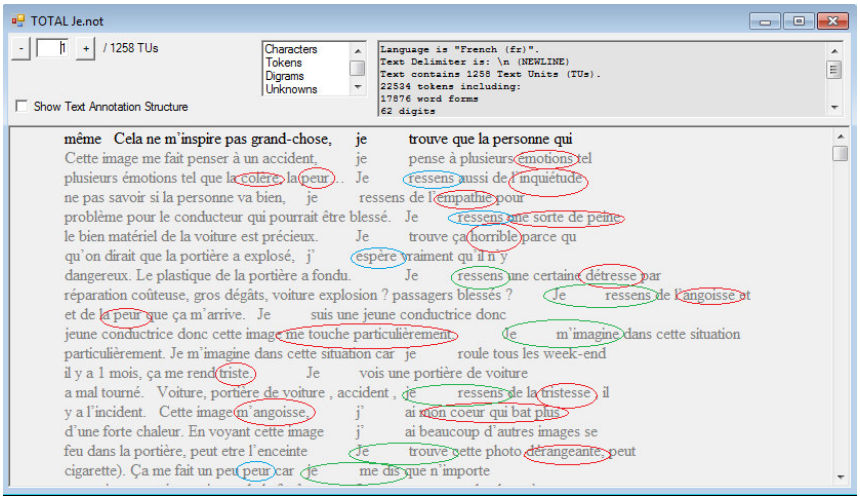


Figure 2. Markers of self-involment in discourse

What is particularly remarkable is the presence of markers of emotion, specific lexicon and verbs. All this has allowed us to see, over 3 years of experience, that emotions are the levers of permeability to infox by image among young people, whether the emotion felt is positive or negative.

4. Conclusion

The addition of semantic information allows the automatic extraction of data that goes beyond the simple lexical form. This allows the search for information based on feelings for example, as we have seen, and will eventually help to identify new levers to fight against the circulation of infoxes. The automation of the recognition of feelings allows, by identifying them precisely, to better understand the mechanisms of reception of these infoxes, but also the mechanisms of their production, in order to better thwart them later.

References

Amossy R. (2010). *La présentation de soi. Ethos et identité verbale*, Paris, Presses universitaires de France.

Bigey M. (2018). Twitter et l'inscription de soi dans le discours. L'ethos pris au piège (ou pas) de la frontière sphère privée/sphère publique , *Les Cahiers du numérique*, 14, p. 55 – 75.

Bigey M. and Simon J. (2020). « Désinfoxiquer les images sur les réseaux socionumériques. Vers une démarche empirique d'éducation à l'image », in Bonfils Ph., Dumas Ph., Remond É. et Stassin B. (dirs), *L'éducation aux médias tout au long de la vie : des nouveaux enjeux pédagogiques à l'accompagnement du citoyen*, Actes du Colloque International TICEMED 12, 7-9 Avril 2020, Athènes, Grèce (p. 198-206). Association Internationale Ticemed, 2021, 978-2-492969-00-3. halshs-03206274V2

Gross M. (1975). *Méthodes en syntaxe : Régimes des constructions complétives*, Hermann, Paris, France.

Mercier A. (2018). dir., *Fake news et post-vérité : tous une part de responsabilité ! , Fake news et post-vérité*, E-book publié par le site The Conversation, p. 4 – 8.

Montaclair F. (2018). *Grammaire de l'infox : une étude de la stylistique de l'information trompeuse*, Condéfence Infodémie, 20 janvier 2022, Besançon.

Silberztein M. (2015). *La formalisation des langues: l'approche de NooJ*, ISTE, Londres.

EMOtivo: a classifier for emotion detection of Italian texts trained on a self-labelled corpus

Marco Bolpagni, Marco Broglio, Andrea Innocenzi, Tommaso Ulivieri

University of Rome “Tor Vergata” -
marco.bolpagni@gmail.com marco.broglio88@gmail.com
andreainnocenzi88@gmail.com tommaso.ulivieri@gmail.com

Abstract
Based on the new techniques implemented in the Natural Language Processing (NLP) field, we here present an emotion detection classifier for Italian texts, built from a corpus generated with texts downloaded from Twitter and self-labelled through the contained emoji. The algorithm was developed by employing the following methodological steps: 1) survey-based study of the association between emoji and emotions in the Italian context; 2) determination of the significant emoji by emotional categories following Ekman's model with the addition of the category 'love'; 3) creation of the corpus by downloading tweets in Italian containing at least one of the selected emoji; 4) cleaning and balancing of the dataset; 5) definition of the model based on the CNN+LSTM hybrid architecture. Using a test set composed of 525 real tweets hand-annotated, the classifier achieved an overall accuracy of 78.9%, correctly classifying 414 tweets.

Keywords: emotion detection, natural language processing, italian text, emoji, self-labelled corpus, deep learning

1. Introduction

The spread of social media in the last decade has supported the adoption of new techniques, such as Natural Language Processing (NLP), to study digital communities. Researchers first focused on the study of sentiment (positive-negative) and, only later on, they shifted their attention towards a more complex approach regarding text interpretation by introducing emotions in computational models (emotion detection) (Medhat *et al.*, 2014). However, these models developed by academics were directed to English texts, being English the most widely used language within Internet communication. Consequently, few scholars have developed models for local languages, such as Italian, which received less attention by researchers (Herring, 2008). Classifiers developed so far for Italian language are not much reliable in real contexts, due to the extremely limited availability, both in quantity and size, of annotated corpora for model training (which are crucial for the model performance).

Given the above, our work aims to develop an emotion detection classifier for texts in Italian language from a sufficiently large self-labelled corpus of downloaded texts from Twitter. As numerous studies indicate a relationship between the emoji used and the overall emotional content of the message, we propose a self-labelling process that exploits emoji contained in tweets to assign the emotional label. As a first step, we created a theoretical framework for linking emoji, emotions and language, on the basis of scientific evidence from various fields of expertise (psychology, anthropology, linguistics).

2. Methodology

Firstly, we conducted a study on the relationship between emoji and emotion, a necessary stage for the identification of emoji to be used for downloading and tagging tweets. After downloading the dataset, we focused on the cleaning and balancing of the dataset and lastly to the training of the model and the evaluation of its performance. The details of the various steps that led to the development of the final classifier are described below.

2.1. Survey and selection of emoji for the various emotions

In the first phase of the project, following a survey, we collected data and information on the association between emoji and emotions in the Italian context. The decision to investigate the emotional connotation of emoji based on a survey came by the absence of studies on the emotional meaning of emoji in the Italian context (Bai *et al*., 2019).Concomitantly, this gave us the advantage to build a sample of emoji based on their effective use and their prevailing emotional meaning without the need to rely on lists of emoji presumptively chosen a priori or research carried out in cultural contexts distant from the Italian one. We based the choice of the emotions to analyze on the model of basic emotions proposed by Ekman, adding love[1].

Data collection took place online from February 26th to March 28th, 2021, using convenience sampling and finalized in a dataset of 1,318 valid responses. The survey was constructed using open-ended, free-entry questions in which the respondent was asked to indicate the most used emoji (maximum 5) for each of the emotions presented.

The resulting dataset was cleaned of any missing or incomplete responses and a total of 276 unique emoji were then identified. From the selected emoji, those with an emotional ambiguity were removed in order to select only emoji with a clear and unambiguous connotation[2]. The remaining emoji were selected on the

[1] Love could be an interesting emotional nuance in areas of application of sentiment analysis such as in marketing, to measure the degree of emotional connection with the brand or product.

[2] To assess emotional ambiguity, an indicator called 'uniqueness' was built and used to filter the set of emoji. This indicator was defined as: $U_{emoji} = \max\left\{\frac{n_{emoji_{emozione}}}{n_{emoji}} : \text{emozione} \in E\right\}$

basis of their frequency distribution by applying the elbow method used in cluster analysis. The selected emoji are summarized in Table 1.

Table 1. Selected set of emoji by emotion

Emotion	Emoji set
Angry	😡🤬😤😠
Sadness	😭😔😢😞🥺☹️
Fear	😰😨
Disgust	🤮🤢
Surprise	😳😲😯😚
Happiness	🤩😊😄🥳😂😁😃☺️🤗😀🤣🎉
Love	😍❤️🥰😘

2.2. The dataset: building, cleaning and balancing

To build the dataset, we downloaded tweets using the API provided by Twitter. All tweets in Italian language published from 1/1/2017 to 28/3/2021 and containing at least one of the emoji selected through the survey were downloaded. To reduce noise, we removed replies and retweets, thus obtaining a corpus of 21.352.942 tweets.

Neutral tweets were also added to the corpus, as the inclusion of a neutral class improves the overall performance of the classifier (Koppel and Schler, 2006). In particular neutral tweets were obtained by scraping messages published by the main accounts of Agenzia Nazionale Stampa Associata (ANSA)[3] in the same time window. Finally, to remove any non-neutral tweets, including quotes or public statements reported in the news, we applied a sentiment analysis algorithm. Overall, the clean neutral subcorpus consisted of 242.026 tweets.

The full dataset, made up of the two subcorpuses, consists of 21,594,968 tweets. Given the strong presence of noise in the corpora generated with texts extracted from Twitter (Xiang *et al*., 2012), we structured a cleaning algorithm as follows:

- bots and language detection: removing any bot users or tweets in languages other than Italian;
- cleaning by tweet attributes: removing tweets with media, links, external texts, duplicate tweets, tweets with more than four hashtags, tweets with references to chains or applications and excessively short tweets;

[3] Considered to be unbiased by mission and communicative tone.

- cleaning by emoji attributes: removing tweets with emoji not identified in the survey, ambiguous emoji and emoji associated with more than one emotion. Following a qualitative analysis on the dataset, some of the emoji associated with the emotional category 'joy' were removed as they were vehicles for ironic or sarcastic messages (Bai *et al*., 2019)
- cleaning up incoherencies: external resources from previous studies were employed to identify discrepancies between the emotional content of the text and that conveyed by the emoji. In this phase, positive-negative sentiment analysis models and the emotional lexicon developed by Passaro and collaborators (2015) were employed.

The cleaned dataset contains 1,192,510 tweets. At the end of the cleaning phase, in order to proceed with the training, we generated an emotive label based on the correspondence between the emoji contained in the tweets and those identified by the survey.

As the distribution analysis of tweets shows an imbalance in the dataset that is worth the attention[4]; we therefore decided to implement a hybrid method to rebalance it that, on the basis of the median value of the distribution of classes (66,932), provides a reduction over-represented classes through a random extraction of examples (random sample) and the generation of artificial texts to increase the number of tweets in under-represented classes. In detail, for the generation of synthetic texts, we implemented a Markovian generative model , which proved better performance (Akkaradamrongrat *et al*., 2019) than other common techniques.

After obtaining a balanced dataset, a data pre-processing step was implemented, aiming to remove irrelevant information for classification and to normalise the text in order to obtain key terms for the purposes of analysis.

The preprocessing steps applied are the following: (1) removal of residual html code present in tweets; turn text into lower case; (2) removal of links, hashtags, residual mentions, numbers and emoji; (3) removal of punctuation except for punctuation with an emotional-intonative function (Encyclopedia Treccani, 2012): "? ", "!", "..."; (4) elimination of stopwords through the customisation of the stopwords-iso list downloaded from *Github*[5]; (5) normalisation of text through the transformation of anomalous forms of expression (such as 'Siiiiiiiiii') into their basic form and stemming of terms.

At the end of the pre-processing phase, we performed a qualitative assessment of the tweets contained in each emotion by taking a random sample of 500 tweets

[4] Several studies suggest that unbalanced distributions of the dataset can considerably affect the performance of the classifier particularly with respect to minority classes (He and Ma, 2013).
[5] Downloaded on 28 March 2021 (https://github.com/stopwords-iso/stopwords-it).

for each emotional category. In this way we were able to verify the quality of the data before proceeding with the training stage of the model. This analysis revealed the presence of highly inconsistent content within the emotion category 'love'. Indeed, tweets associated with this emotion label were found to be confusing and overlapping with other emotions. We eventually decided to remove the 'love' category and step back to the emotional classification based on the categories identified by Ekman *et al*. (1972): sadness, anger, surprise, happiness, fear and disgust. Hence, the final corpus used for training the model was composed of 468,524 tweets.

2.3. Model development and training

For the development of the classifier's architecture we referred tothe state of the art models for emotion detection in text. The literature pointed to LSTM (Long Short Term Memory) layers that allow the extraction of features considering the context and to the position of terms in the sentence. Nevertheless these are not very effective in the management of short term relationships between terms such as in the case of idioms and idiomatic expressions (Ombabi *et al*., 2020). To overcome the limitations of architectures composed only of LSTM layers, we decided to apply a hybrid model with the addition of a convolutional layer (Convolutional Neural Network or CNN) which is very effective in extracting local textual features (Kim, 2014). In the final architecture, an upstream embedding layer and a max pooling layer between the CNN and the LSTM has been added[6]. Finally, a dense layer with a soft-max activation function was added to handle the actual classification.

Once the framework was defined, we established the size of the dictionary to be used for building the classifier using Zipf's law. Based on the distribution of terms contained in the tweets, the dictionary used for our classifier is composed of 16,886 terms, representing the top 20% of the terms contained in the training corpus sorted by frequency.

The subsequent step in the definition of the layers that make up the model architecture was the identification of the hyperparameters. This was carried out through a Bayesian cross-validation search (k-fold = 4) and considering accuracy as a metric for evaluating the models. Specifically, the investigation concerned the following hyper-parameters : the size of the embedding (optimal value = 136), the number and the size of the kernels of the convolutional layer (optimal values = 136, 4), the activation function of the convolutional layer (optimal value = ReLU), the size of the pool considered by the max pooling layer (optimal value = 4), the number of neurons of the LSTM layer (optimal value = 56) and the

[6] In order to reduce the processing time and the risk of overfitting.

optimisation algorithm (optimal value = Adam).
After identifying the optimal parameters, the model was trained with TensorFlow 2.4.0 on Python 3.8.5. For the training of the model the dataset was divided into training set and validation set, according to an 80-20 scheme and a batch size of 16 was set. We considered categorical cross entropy as loss function and set the maximum number of training epochs at 15.. In order to avoid overfitting, early stopping was introduced, interrupting training and re-establishing the best weights after 3 epochs without significant improvement.

3. Performance evaluation

3.1. Test sets and rating metrics

For the evaluation of our classifier we employed a human-annotated test set[7] consisting of 525 real tweets published in a period following the scraping of the training set. The tweets in the test set were selected by means of search keys from four different annotators and tagged on the basis of the main emotion. At the end of the tagging phase, the quality of the data was checked by selecting only tweets with a clear emotional connotation shared by all annotators. Finally, in order to standardise the contents of the test set, the tweets were preprocessed by replicating the algorithm used for the training corpus.
Precision, recall and F1-score on one side, and accuracy on the other were considered for the evaluation of the performance related to individual emotions and the overall performanceof the classifier, respectively.

3.2. Outcomes

The classifier achieved an overall accuracy of 78.9%, correctly classifying 414 tweets out of the 525 contained in the test set. Table 2 shows the classifier's precision, recall and F1-Score metrics for each of the emotions considered.

Table 2. Model performance for any emotion

Emotion	Precision	Recall	F1-score
Disgust	0,84	0,85	0,85
Happiness	0,77	0,85	0,81
Neutral	0,89	0,84	0,86
Fear	0,82	0,89	0,85
Anger	0,70	0,77	0,73
Surprise	0,86	0,57	0,69
Sadness	0,69	0,73	0,71

[7] Using self-labelled datasets adds noise and bias that affect performance assessment (Kaity and

4. Conclusions

In this work, we built an emotion detection classifier from a corpus composed of texts downloaded from Twitter and self-labelled through the emoji contained in them. The performance of the model shows that it is possible to train reliable classifiers for emotion detection from self-labelled corpora by using emoji. It also shows that tweets can convey several emotions simultaneously in a real context. Therefore, it would be interesting to re-elaborate the work using a multi-label approach more suitable for capturing the coexistence of multiple emotional nuances. The proposed architecture, CNN+LSTM, proved to be able to capture the emotional meaning of many of the idiomatic expressions present in the test set, confirming it as a good solution for emotion detection problems. The quality of the classifier, however, is closely linked to the attention paid to the cleaning and pre-processing stages, which are fundamental for improving the quality of the input data.

References

Akkaradamrongrat, S., Kachamas, P., and Sinthupinyo, S. (2019). Text Generation for Imbalanced Text Classification. In *2019 16th International Joint Conference on Computer Science and Software Engineering* (JCSSE), pp. 181-186.

Bai, Q., Dan, Q., Mu, Z., and Yang, M. (2019). A Systematic Review of Emoji: Current Research and Future Perspectives. *Frontiers in psychology*, vol.10: 2221.

Ekman, P., Friesen, W. V., and Ellsworth, P. (1972). *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Pergamon.

He, H., and Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press.

Herring, S. C. (2008). *Language and the Internet*. Cambridge University Press.

Kaity, M., and Balakrishnan, V. (2020). Sentiment lexicons and non-English languages: a survey. *Knowledge and Information Systems*: 1-36.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 1746–1751.

Koppel, M., and Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, vol.22(2): 100-109.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, vol.5(4): 1093-1113.

Ombabi, A. H., Ouarda, W., and Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, vol.10(1): 1-13.

Passaro, L., Pollacci, L., and Lenci, A. (2015). ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon. In *Second Italian Conference on Computational Linguistics CLiC-it*, pp. 215-220.

Xiang, G., Fan, B., Wang, L., Hong, J. I., and Rose, C. P. (2012). Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In *CIKM '12: Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1980-1984.

“Altro, specificare”. Effetto del Covid-19 sulla codifica dei campi “altro” nell’indagine Istat sulle forze di lavoro[1]

Barbara Boschetto[1], Francesca della Ratta[2], Federica Pintaldi[1], Maria Elena Pontecorvo[1], Alessia Sabbatini[1]

[1]ISTAT – boschett@istat.it pintaldi@istat.it mariaelena.pontecorvo@istat.it sabbatini@istat.it
[2]INAPP – f.dellaratta@inapp.org

Abstract
With the outbreak of the pandemic, there was a strongly demand to provide information about the exceptional situation in the country. For current surveys, it was not possible to revise the questionnaires to measure the impact of the pandemic on phenomena under investigation. In the Labour Force Survey an added value is provided by the "other specify" fields in which the interviewers had the possibility to record all the answers related to the emergency situation. The set of open answers inserted in the "other" fields was analysed with the aim to extract for each respondent the answers connected to Covid-19. In this way a corpus of 110,859 occurrences and 5.567 different forms was analysed, using as a starting point the selection of 17,995 records which mentioned a Covid-related answer in the "other" field. The set of selected answers offers an interesting insight on the issues that have most impacted on the labour market in the first year of the pandemic.

Keywords: Text Mining; Covid19; Labour Force Survey; Istat; data integration

Riassunto
Con lo scoppio della pandemia è emersa la richiesta, nei confronti della statistica pubblica, di dar conto della situazione eccezionale che si è verificata nel Paese. In questo contesto le informazioni raccolte dalla Rilevazione sulle forze di lavoro sono state valorizzate attraverso l’analisi dei campi aperti presenti nel questionario ricchi di riferimenti all’emergenza sanitaria che non sarebbero emersi altrimenti, poiché non incasellabili negli item pre-codificati.

Parole Chiave: Text Mining; Covid19; Labour Force Survey; Istat; data integration

[1] Il lavoro è frutto di un lavoro comune. Le opinioni espresse sono da riferire alle autrici e non agli istituti di appartenenza. Premessa e conclusioni sono state redatte da tutte le autrici, il paragrafo 2 da Federica Pintaldi, il 2.1 da Alessia Sabatini, il 2.2 da Barbara Boschetto, il 3 da Francesca della Ratta e il 4 da Maria Elena Pontecorvo.

1. Premessa

Con lo scoppio della pandemia è emersa la richiesta, nei confronti della statistica pubblica, di dar conto della situazione eccezionale che si è verificata nel Paese. Tuttavia non sempre le rilevazioni hanno avuto modo di intervenire sui questionari per cogliere l'impatto della pandemia sui fenomeni indagati. Nel caso della Rilevazione Istat sulle forze di lavoro un valore aggiunto è fornito dai campi "altro, specificare", in cui i rilevatori hanno potuto annotare le risposte connesse alla situazione emergenziale non incasellabili negli item previsti dal questionario. L'insieme dei testi inseriti nei campi "altro" è stato analizzato per estrarre in modo automatico dal dataset dell'indagine (circa 483 mila interviste a persone con più di 15 anni nel corso del 2020) il sottoinsieme di record in cui era presente almeno un riferimento alla pandemia. Alle risposte "Covid" individuate automaticamente è stata quindi attribuita la codifica "altro-Covid", da aggiungere al file in analisi, offrendo un interessante esempio di integrazione tra tecniche di analisi e tipi di dati.

2. La costruzione del corpus

Nel corso dell'emergenza è stato indicato ai rilevatori di inserire nelle risposte "altro" le informazioni che, a causa della pandemia, non riuscivano a incasellare negli item predefiniti del questionario. L'insieme dei campi "altro" del primo trimestre 2020 è stato sottoposto ad analisi testuale[2] per estrarre una prima lista di espressioni riconducibili al Covid. Questa lista, composta da circa 40 forme grafiche, è stata utilizzata in ciascuno dei quattro trimestri del 2020 per estrarre automaticamente tutti i record con almeno un riferimento al Covid-19 nei campi "altro, specificare". Ai record così estratti, successivamente validati dai ricercatori, è stata attribuita la codifica specifica "altro-Covid". Si tratta nel complesso di 17.995 record, il 61% dei quali riferiti al secondo trimestre 2020[3].

2.1. Le risposte aperte nel questionario core

Tra i quesiti presenti nel questionario, quelli con la quota maggiore di specifiche "altro-Covid" sono riferiti ai motivi: dell'assenza dal lavoro, della variazione nell'orario di lavoro, della mancata ricerca di lavoro e dell'interruzione

[2] L'analisi è stata condotta con TaltaC2 (**Errore. Riferimento a collegamento ipertestuale non valido.**), per la nuvola di punti è stato utilizzato IRaMuTeQ (http://www.iramuteq.org/)

[3] Il ricorso alle risposte "Altro" è stato particolarmente consistente nel 2020. Ad esempio nella domanda B3 sulle assenze nella media dell'anno sono state registrate il 20,4% di risposte altro, a fronte dell'1,3% del 2019; nella domanda F10 la quota di "Altro" è passata dal 1,0% del 2019 al 2,7% del 2020, con un picco del 5,1% nel secondo trimestre. Data la natura prettamente metodologica del contributo, si tratta di dati campionari senza riporto all'universo.

dell'attività lavorativa[4]. Nella tabella 1 sono riportate le percentuali della specifica "altro-Covid" sul totale delle risposte "altro" e sul totale dei rispondenti.

Tabella 1. Quesiti in cui nei campi aperti sono presenti riferimenti al Covid – Anno 2020

Motivi:	% Covid su Altro	% Covid sul Totale	N rispondenti al quesito
dell'assenza dal lavoro (B3)	78,6	16,1	27.106
per cui ha lavorato più del solito (C33)	58,9	6,3	4.954
per cui ha lavorato meno del solito (C34)	10,4	1,4	45.024
per cui non sarebbe disponibile a lavorare più ore se ne avesse l'opportunità (C41)	76,0	44,0	2.232
per cui ha smesso di lavorare (solo dipendenti, E14)	24,7	1,0	52.882
per cui ha smesso di lavorare (solo autonomi, E15)	40,3	2,4	1.162
per cui non ha cercato un lavoro nelle 4 settimane precedenti l'intervista (F10)	59,5	1,6	189.624
per cui non avrebbe potuto iniziare a lavorare (no ricerca 4 settimane) (F14)	71,0	9,7	7.551
per cui non avrebbe potuto iniziare a lavorare (ricerca nelle 4 settimane) (F34)	71,8	8,5	2.078

Tra i quesiti considerati, quello con il peso maggiore di risposte Covid è quello sul motivo dell'assenza dal lavoro nella settimana di riferimento (78,6% sul totale delle risposte "altro"): con 4.351 risposte riferite al Covid su un totale di 27.106 rispondenti al quesito. Altra domanda con un peso consistente della specifica Covid è quella sulla mancata ricerca nelle 4 settimane precedenti l'intervista: infatti, una delle conseguenze del Covid sul mercato del lavoro, oltre al calo nel numero di occupati, è stata anche la riduzione di persone in cerca di lavoro (disoccupati), un dato all'apparenza anomalo perché a un forte calo nell'occupazione generalmente corrisponde un aumento della disoccupazione. Tuttavia, nel contesto pandemico non era facile cercare attivamente un lavoro (ad esempio a causa della chiusura delle attività o delle limitazioni imposte alla mobilità).

2.2. Le risposte aperte nel modulo ad hoc

L'evoluzione della crisi sanitaria emerge anche dall'uso delle risposte aperte nei quesiti del modulo ad hoc "Salute e sicurezza sul lavoro", approfondimento tematico inserito nell'anno 2020 (Istat, 2021), in particolare per il quesito sulla percezione di fattori di rischio per la salute fisica sul posto di lavoro. Per indagare questo aspetto era proposta agli intervistati una lista di dieci tipi di rischio più una risposta aperta "altro, specificare". Il 62,4% degli occupati ha dichiarato di sentirsi esposto ad almeno un fattore di rischio per la salute fisica e l'8,9% ha utilizzato la risposta "altro". Tra questi il 18,8% ha menzionato esplicitamente il

[4] Tra i quesiti aperti presi in considerazione vi è anche il campo finale a cura del rilevatore in cui spesso sono riportate osservazioni sull'intervista. Il campo è tuttavia riferito non al singolo individuo ma a tutta la famiglia, motivo per cui non viene riportato nella tabella che riporta i casi individuali.

Covid (l'1,7% degli occupati), e un ulteriore 47,7% ha indicato un più generico rischio di infezione (il 4,3% degli occupati) utilizzando parole come *rischio di infezione, contagio, virus, pazienti*, o termini meno inerenti la pandemia, come *batteri* o *laboratorio*. La percezione del Covid come fattore di rischio sul lavoro è cresciuta nel corso dei mesi, passando dallo 0,4% nel primo trimestre 2020 all'1,4% del secondo trimestre, fino al picco del 3,0% nel quarto trimestre 2020. La percezione del rischio Covid è diffusa soprattutto fra gli occupati delle professioni intellettuali e scientifiche (che include medici, farmacisti, professori e insegnanti) fra i quali la percentuale sale al 2,8% (più il 7,6% del rischio generico di infezioni), seguiti dalle professioni tecniche (fra cui infermieri e tecnici di laboratorio) che hanno indicato il rischio Covid nel 2,2% dei casi (più il 7,1% che segnala il rischio generico di infezioni) e dalle professioni qualificate nelle attività del commercio e dei servizi, che hanno indicato il Covid nel 2,3% dei casi (più 4,9% per le infezioni).

3. Le caratteristiche del corpus

Nel complesso, il corpus composto dai testi delle stringhe del campo "altro, specificare" contenuti in tutti i record in cui è presente almeno un riferimento al Covid, è composto da 110.859 occorrenze e 5.567 forme diverse, con un

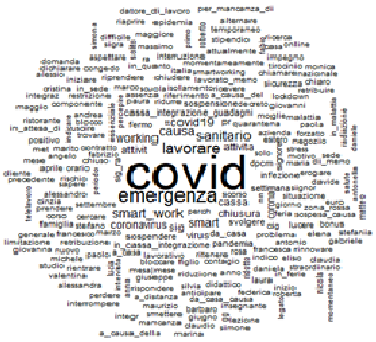


Figura 1. Nuvola di punti – Corpus "altro,specificare" ottenuto con Iramuteq

indicatore di ricchezza lessicale (V/N) piuttosto basso (5), tipico di frammenti molto brevi. Tra le parole e le polirematiche che si incontrano con maggiore frequenza vi sono: *covid, emergenza sanitaria, smart work, coronavirus, lavoro, attività, chiusura, cassa integrazione guadagni, da casa, sospesa, pandemia, dpcm e quarantena* (Figura 1). Tra i segmenti ripetuti, oltre alle diverse espressioni utilizzate per indicare l'emergenza sanitaria e la condizione di eventuale contagio o isolamento (*emergenza Covid19, contagio Covid, pandemia Covid19, zona rossa/gialla, quarantena coronavirus*, ecc., per un totale di circa 13 mila occorrenze considerando i segmenti con più di 3 occorrenze), si incontrano numerosi riferimenti al lavoro da remoto (*smart work, smartworking*

per Covid, lavorando in smart working, entrambi smartworking, lavoro agile, ecc, per un totale di 4.713 occorrenze); alla chiusura delle attività non essenziali e al ricorso alla cassa integrazione (*non essenziale, meno ore, dpcm Covid19, momentaneamente sospesa, in cassa integrazione per Covid, ritenuta non essenziale*, 2.895 occorrenze); alla chiusura delle scuole o la didattica a distanza (*chiusura scuola, didattica a distanza, didattica online, video lezioni, collaboratrice scolastica*, 192 occorrenze) e all'impossibilità di cercare lavoro nel contesto pandemico (*non cerca lavoro, vorrebbe lavorare, assunzioni bloccate*, 136 occorrenze). Un'ulteriore insieme di parole significative è quello costituito dalle cosiddette parole chiave, vale a dire quelle la cui frequenza relativa nel testo in esame è superiore a quella con cui le stesse parole compaiono in un lessico di frequenza assunto come riferimento[5]. Tra quelle con lo scarto più elevato vi sono tutte quelle che rimandano alla situazione di pandemia e alla chiusura temporanea di alcune attività: *emergenza, infezioni, lavorato virus, sospesa, chiusura, corona, cassa, sospensione, contagio, isolamento, ridotta, restrizioni, ferie, forzate, sospeso, stress, bonus, bloccato, epidemia*. Grazie all'analisi lessicale fin qui illustrata è stato possibile quindi individuare i principali nodi tematici presenti nelle risposte: 1) l'esperienza del lavoro da remoto, 2) la chiusura delle attività e il ricorso massiccio alla cassa integrazione, 3) la chiusura delle scuole e l'esperienza di isolamento o quarantena. Il passaggio all'analisi testuale (Bolasco, 2013) ha consentito di ottenere una quantificazione di ciascuno di questi nodi all'interno del corpus, rintracciando (sempre utilizzando la funzione di Ricerca Entità) nel testo tutti i termini riferiti a ciascuno di essi e aggiungendo cinque nuove variabili al dataset di partenza, popolate quando uno dei frammenti del corpus conteneva un riferimento a uno dei nodi tematici considerati. Il dataset di partenza quindi si è arricchito così di nuove variabili, una tecnica molto utile nel caso di domande aperte (della Ratta, 2010). Nella tabella 2, in cui sono riportati i risultati di questa operazione, appare evidente la preminenza del riferimento al lavoro da remoto nelle risposte degli intervistati, insieme al tema della chiusura delle attività e della cassa integrazione, soprattutto nei mesi di aprile e maggio 2020. In particolare, i cenni alla cassa integrazione, citati spesso nella domanda finale sulle osservazioni dei rilevatori in cui erano sintetizzate le condizioni emerse nel corso dell'intervista, possono essere suddivisi tra quelli riportati dalla maggioranza di lavoratori messi in cassa integrazione e una minoranza che invece ha segnalato di non aver potuto usufruire di questa forma di sostegno al reddito. Meno numerosi i riferimenti

[5] Per estrarre le parole chiave è stato utilizzato come riferimento il lessico dell'italiano standard, contenuto nel software Taltac2.

alla chiusura delle scuole, concentrati soprattutto a inizio pandemia, e allo stato di quarantena[6].
La possibilità di disporre di variabili testuali consente inoltre di individuare relazioni con le variabili del questionario: ad esempio il riferimento allo *smartworking* è più frequente tra coloro che svolgono una professione compatibile con il lavoro da remoto (della Ratta, Gallo, Sabbatini; 2021), mentre i riferimenti alla cassa integrazione sono diffusi soprattutto tra chi è occupato in una professione operaia o nel settore dei servizi, così come le chiusure sono menzionate soprattutto da quanti svolgono una professione nel settore dei servizi o tra gli imprenditori.

Tabella 2. Riferimenti ad alcune categorie tematiche per mese (valori assoluti e % su totale record)

	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic	**totale**	% sui record
SMARTWORKING			31	793	1.197	292	191	70	87	149	190	167	**3.167**	17,6
CHIUSURA	3	18	253	701	444	69	23	12	12	44	122	74	**1.775**	9,9
di cui:														
per attività non essenziale			16	117	73	3			1		1	3	**214**	1,2
CASSA INTEGRAZIONE		1	20	632	645	120	31	19	15	29	38	31	**1.581**	8,8
di cui:														
è stato messo in Cig		1	18	576	611	114	31	19	14	29	38	31	**1.482**	8,2
non è stato messo in Cig			2	56	34	6			1				**99**	0,6
SCUOLA		4	37	64	99	29	7	2	7	8	9	6	**272**	1,5
QUARANTENA		16	85	22	17	3	2	1	8	20	33	14	**221**	1,2

4. Parole caratteristiche

La relazione tra le variabili del questionario e il contenuto testuale del corpus costituisce sicuramente uno degli aspetti più interessanti dell'integrazione tra dato quantitativo e dato testuale. Come noto, utilizzando le variabili associate alle risposte è possibile rintracciare alcune differenze nel lessico utilizzato, individuando alcuni gruppi di parole (dette parole caratteristiche), che, per la loro sovra-rappresentazione in specifiche partizioni del testo, consentono di caratterizzarne lo specifico profilo lessicale (Lebart L., Salem A., 1994; Tuzzi, 2003). Tra i risultati più interessanti vi sono le specificità per mese di intervista, che riproducono l'evoluzione della pandemia con un linguaggio via via più specifico e indicativo della configurazione che la pandemia andava assumendo

[6] Il tema delle difficoltà nella ricerca di lavoro risulta numericamente meno rilevante per via del minore peso relativo degli individui che non cercano lavoro per via del Covid rispetto al peso degli occupati. Lo stesso si può dire per le domande del modulo ad Hoc somministrate solo a un quarto del campione.

mese per mese (Figura 2). Ecco dunque che se all'inizio dell'anno risultano caratteristici termini come *rischio*, *isolamento* e *quarantena* a marzo esplodono i riferimenti all'*emergenza sanitaria*, con l'aggiunta, a partire da aprile, dei termini *essenziale*, *sospensione*, *dpcm*, *cassa integrazione*, e *smartworking*.

GEN	FEB	MAR	APR	MAG	GIU	LUG	AGO	SET	OTT	NOV	DIC
chiusura	coronavirus	emergenza sanitaria		smartworking			infezioni	smartworking	infezioni	zona rossa	pandemia
aziendale	isolamento	19	cassa integrazione				stress	possibile	paura	infezioni	zona rossa
rischio	quarantena	coronavirus	non	dpcm			paura	contagio	quarantena	positivo	contagio
		virus	essenziale					infezioni		lockdown	covid
		corona	sospensione					paura			smartworking
		covid19	sospesa								

Figura 2. Parole specifiche per mese

Queste ultime, sono peraltro le uniche parole specifiche dei mesi estivi che hanno visto l'allentarsi delle restrizioni e la temporanea sensazione che l'emergenza stesse volgendo al termine. Fanno eccezione i riferimenti a *infezioni* e *stress* nel mese di agosto, come già visto specifiche delle domande del modulo ad hoc inerenti alla percezione dei fattori di rischio fisico e psicologico connesse al lavoro. Nei mesi autunnali, il riacutizzarsi dei contagi e il ritorno delle misure restrittive si manifestano con parole come *paura*, *quarantena*, *lockdown* e *zona rossa*.
Altrettanto interessanti sono le specificità per professione svolta che accomunano di fronte all'emergenza sanitaria alcuni gruppi professionali anche lontani tra loro. Ad esempio, le parole *attività sospesa*, *essenziale*, o *ridotta attività*, sono specifiche sia del gruppo che comprende dirigenti e imprenditori, sia di quello che comprende, tra gli altri, artigiani e agricoltori, ovvero le professioni in cui è più elevata la quota di lavoratori autonomi, che hanno più sofferto della sospensione delle proprie attività. Parole che si riferiscono all'ambito scolastico come *didattica*, *distanza*, *da casa*, ma anche *scuola*, sono specifiche del gruppo II (professioni intellettuali e scientifiche), che naturalmente include professori e insegnanti, e anche del gruppo VIII (professioni non qualificate) che comprende bidelli, custodi o personale addetto alla pulizia che lavorano presso le scuole che erano chiuse. Il riferimento allo *smartworking* è specifico oltre che per il gruppo II anche per il gruppo III (professioni tecniche) e IV (esecutive nel lavoro di ufficio). Un gruppo molto specifico nelle parole che lo caratterizzano è il gruppo V (professioni qualificate nelle attività del commercio e dei servizi) che include professioni specifiche di settori, come il commercio o la ristorazione, che più di altri sono stati interessati dalle restrizioni imposte dai Dpcm: spiccano infatti termini come *negozio*, *essenziale*, *zona rossa*, ma anche *asporto*, *decreto* e *momentaneamente*. I riferimenti alla *cassa integrazione* sono invece specifici del gruppo VII (conduttori di impianti) in quanto strumento molto utilizzato nell'industria.

5. Conclusioni

L'analisi fin qui svolta si colloca nell'ambito di una riflessione sulle potenzialità offerte dall'integrazione tra tipi di dato e sul ruolo delle domande aperte anche in campioni molto vasti come quello della Rilevazione sulle forze di lavoro. Peraltro, oltre al valore conoscitivo che può essere offerto dall'analisi dei campi aperti per fornire spiegazioni aggiuntive a particolari fenomeni, come nel caso del calo della disoccupazione e aumento dell'inattività nei primi trimestri di pandemia, l'analisi dei campi "altro" ricopre un ruolo rilevante anche nelle operazioni di miglioramento continuo dei dati. Infatti, nel corso del 2020, il monitoraggio dei campi aperti nella situazione eccezionale di pandemia ha consentito di fornire ai rilevatori indicazioni rispetto a eventuali errori o imprecisioni nella codifica: l'attività di restituzione ai rilevatori è infatti una delle ragioni della riduzione dell'impiego della modalità "altro" nell'ultima parte dell'anno. Tale attività peraltro non è una novità per la Rilevazione sulle forze di lavoro, che da anni porta avanti un processo di miglioramento della codifica del campo relativo all'attività economica proprio grazie a tecniche di analisi testuale (della Ratta, Tibaldi; 2014).

Bibliografia

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Carocci, Roma.

della Ratta-Rinaldi F. and M. Tibaldi (2014). "Sperimentazione di un sistema di controllo e correzione per la codifica dell'attività economica". *Istat Working Paper*, n. 4, 2014.

della Ratta-Rinaldi F. (2010). "Se pensa al suo futuro, di cosa ha più paura?", in S. Bolasco, I. Chiari, L. Giuliano (a cura di). *Statistical Analysis of Textual Data. Proceedings of 10th International Conference 9-10 June 2010*.

della Ratta-Rinaldi F., F. Gallo and A. Sabbatini (2021). "Il lavoro da remoto: potenzialità e pratica prima e durante la pandemia da Covid-19". *Rassegna italiana di sociologia*. A. LXII, n. 2, aprile-giugno.

Istat (2021), Salute e sicurezza sul lavoro in Italia, Comunicato stampa, 29 novembre 2021. https://www.istat.it/it/files//2021/11/Salute_e_sicurezza_sul_lavoro_-2020_Istat_Eurostat.pdf

Lebart L. and Salem A. (1994). Statistique textuelle. Paris, Dunod.

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

Évaluation et description de récits oraux spontanés d'élèves du primaire : pour l'utilisation des méthodes mixtes en sciences de l'éducation

Rosalie Bourdages[1,2], Roxane Gagnon[2], Denis Foucambert[1]

[1]Université du Québec à Montréal – rosalie.bourdages@hepl.ch
foucambert.denis@uqam.ca
[2]Haute école pédagogique du canton de Vaud - roxane.gagnon@hepl.ch

Abstract
The teaching of oral language is plagued by difficulties in terms of its assessment (Nonnon, 2016). In order to offer solutions for the measurement of oral language, we have constructed an evaluation grid dedicated to the description of transcribed oral data. Four dimensions of oral production are represented: performative measures, structural measures, linguistic-discursive measures, and narrative length measures. We analyzed 78 spontaneous oral pre- and posttest productions of 46 5- to 10-year-old students, using these dimensions (Griffin *et al*., 2004; Kern, 1997; Schneuwly et Dolz, 1987). Our paper presents the results of a textual analysis of oral discourse conducted with mixed methods. The analyses show longer narratives, without any increase in vocabulary or grammatical complexity. The cross-referencing of narrative and linguistic data highlights underlying dimensions specific to oral narration. Our method allows us to highlight the advantages for educational sciences to move away from the dichotomous positions of the quantitative-qualitative towards a search for theoretical and methodological relevance.

Keywords: Textual data analysis; mixed methods; oral production; narrative; linguistic analysis.

Résumé
L'enseignement de l'oral est marqué par des difficultés sur le plan de son évaluation (Nonnon, 2016). Afin d'offrir des pistes de solutions pour la mesure de l'oral, nous avons construit une grille d'évaluation dédiée à la description de données orales retranscrites. Quatre dimensions de l'oral sont représentées : mesures performatives, mesures structurelles et actants, mesures linguistico-discursives et mesures de longueur de la narration. Nous avons analysé 78 productions orales spontanées d'élèves de 5 à 10 ans à l'aide de ces dimensions constitutives (Griffin *et al*., 2004; Kern, 1997; Schneuwly et Dolz, 1987). Notre communication présente les résultats d'une analyse textuelle de discours oraux réalisée avec des méthodes mixtes. Les analyses font état de récits plus longs, sans que le vocabulaire ne soit plus riche ou les structures grammaticales plus complexes. Les croisements entre données narratives, linguistiques et performatives mettent en exergue des dimensions sous-jacentes inédites propres à l'oral. La méthode adoptée met en valeur les avantages, pour les sciences de l'éducation, à délaisser les

positions dichotomiques du quantitatif-qualitatif pour aller vers la recherche de la pertinence théorique et méthodologique.

Mots-clés : Analyse des données textuelles; méthodes mixtes; oral; narration; analyse linguistique.

1. Introduction

Les difficultés vécues par les didacticien•ne•s et les enseignant•e•s concernant l'enseignement de l'oral s'expliquent en partie par le manque de critères bien définis permettant son évaluation (Nonnon, 2016). La mesure des habiletés à l'oral des élèves du primaire est minée notamment par le recours fréquent à des critères soit très généraux, soit très proches de ceux utilisés à l'écrit (Hassan, 2011). Dans le but de construire un outil reprenant les éléments constitutifs d'une production orale chez des élèves du primaire, nous avons conçu une grille d'évaluation de l'oral narré inspirée des travaux en psycholinguistique (Diez-Itza *et al*., 2017; Drijbooms *et al*., 2017), en acquisition du langage et en didactique (De Weck et Rosat, 2003; Dolz et Schneuwly, 2017 [1998]; Gagnon et Dolz, 2016), qui se déploie en une liste de critères, répartis en quatre ensembles, ou dimensions. À l'aide de cette grille, nous avons analysé les 78 narrations orales de 46 élèves de 5 à 10 ans, prises à deux moments différents : l'un avant la présentation d'une séquence didactique sur le conte, et l'autre après. Trois questions de recherche posent les jalons de nos analyses :

- Peut-on observer des progressions chez les élèves entre la production initiale (PI) et la production finale (PF)?
- Comment se caractérisent ces productions orales de récits spontanés?
- Est-ce que les méthodes mixtes sont à même de répondre à ces questions?

Notre communication présente les résultats préliminaires de cette analyse textuelle de discours oraux réalisée avec des méthodes mixtes et largement exploratoires.

2. Cadre théorico-méthodologique

Le conte oral comme genre textuel fait partie du cursus à l'école primaire. Il permet de travailler le langage d'évocation : les élèves sont amené•es à objectiver, à mettre à distance les événements, les personnages et leurs actions (Bronckart, 1997). En apprenant à raconter une histoire, ils et elles développent des habiletés transférables à la situation d'écriture : la cohérence globale, la cohésion locale, la sélection des contenus la capacité de lier l'énoncé au reste du texte sont utiles à la production et à la compréhension de textes écrits.

Notre grille d'évaluation de la production du discours narratif comporte 4 grandes dimensions constitutives, chacune constituée de plusieurs sous-éléments:

- les mesures de la longueur de la narration ;
    - Nombre de mots total
    - Nombre d'énoncés
- les mesures linguistico-discursives ;
    - Organisateurs textuels spatiaux et temporels
    - Pronoms déictiques et anaphoriques
    - Maintien de la cohésion nominale (chaine des reprises)
    - Temps de verbe utilisés : emploi du temps (PS/IMP/PC/PRES…) et maintien du système de temps (PS/IMP ou PC/IMP ou Prés)
    - Lexique : nombre et richesse de lemmes (fréquence lexicale)
    - Éléments syntaxiques complexes propres au français parlé : doubles instanciations du sujet (redoublement du sujet); dispositifs clivés en « c'est...que /qui »; dispositifs en « il y a SN que/qui » (SN : syntagme nominal).
- les mesures structurelles et actants ;
    - État initial : mention des personnages principaux; mention de leurs actions ou de leurs états mentaux; mention de la localisation spatiale et de la localisation temporelle
    - Élément déclencheur (provocation) : mention de l'élément déclencheur; mention de l'émotion créée par le déclencheur
    - Actions : mention de tentative de résolution du problème dans la suite logique du récit; mention d'un adjuvant
    - Sanction, conséquence : mention de conséquence négative; mention de l'émotion créée par la conséquence
    - État final : lien avec l'état initial; commentaires sur l'émotion engendrée par le retour au calme/état final
- les mesures performatives.

    Prise en compte de la situation de communication : adoption d'une posture de conteur
    - Insertion de dialogues entre les personnages
    - Présence d'une ouverture du récit (appel initial au destinataire)
    - Présence d'une clôture (appel final au destinataire)
    - Gestion des pauses

3. Méthode et analyses

Le corpus oral, d'un total de 11 548 mots, est composé de 78 productions (n=35 en PI, 43 en PF ; n=26 au cycle 1, 56 au cycle 2), transcrites et codées pour les mesures linguistiques et de durée de la narration grâce à l'outil d'analyse lexicométrique automatique AnaText (Kraif, 2012). Le codage de la dimension performative, des structures propres au français parlé et des mesures structurelles et actants a été effectué par l'entremise du logiciel d'analyses qualitatives NVivo (QSR International Pty Ltd, 2020).

D'abord, nous avons soumis nos données prétest et posttest à des analyses quantitatives descriptives et exploratoires permettant la mise en parallèle des résultats aux deux passations, afin, d'une part, de mettre en exergue les éléments ayant évolué entre les productions initiales (désormais PI) et les productions finales (désormais PF) et d'autre part, de dégager les grandes tendances structurelles dans les productions des élèves.
Ces analyses statistiques des données textuelles ont guidé les analyses qualitatives. Les deux types d'analyses utilisées nous permettent de vérifier si la grille atteste de variations interindividuelles et de progrès, signes de son efficacité pour l'enseignement.

3.1. Analyses des mesures de la longueur des narrations

Les analyses t de student pour échantillons appariés, effectuées sur Jamovi (The jamovi project, 2021), font état de récits plus longs (nombre de mots : $t = 3{,}90$, $p<0{,}001$; nombre d'énoncés : $t = -4{,}48$, $p<0{,}001$). Les productions en PI font en moyenne 111,16 mots et 13,81 énoncés. En PF, les récits comportent en moyenne 185,88 mots et 23,09 énoncés. Ces résultats sont le signe de contes spontanés plus fournis en productions finales, et dénotent une confiance accrue chez les élèves post-séquence d'enseignement.

3.2. Analyses des mesures structurelles et actants

Des analyses quantitatives de nature descriptive ont été effectuées sur les données pré et posttest des mesures structurelles et actants. Celles-ci dénotent des formes de progression pour plusieurs composantes et sous-composantes de cette dimension. La mention du personnage principal est généralisée, ce qui induit un effet plafond (à l'instar de Diez-Itza *et al*., 2017), et la présence d'un état initial est fortement majoritaire et stable entre la PI (31 productions) et la PF (32 productions). En ce qui a trait à la présence d'un état final dans les récits, la progression est nette : le nombre de récits incluant un état final passe de 11 en PI à 29 en PF. Onze de ces états finaux, soit environ le tiers des productions, incluent un rappel de l'état initial en PF, contre 3 en PI. La sanction /conséquence est, somme toute, un élément du conte peu mobilisé dans les productions des élèves, mais son nombre double entre le pré et posttest, figurant dans 6 des PI, et dans 12 des PF.
Nos analyses qualitatives des productions permettent de dégager quelques tendances générales. La situation initiale permet de donner une représentation basique de la situation (à l'instar de Godard, 2009, notamment). La localisation temporelle est toutefois absente des productions, hormis quelques exceptions du type « c'était la nuit » ou encore « c'était l'heure d'ouvrir les cadeaux \ ». Une forme de localisation spatiale est présente dans le tiers des productions (ex. « dans l'espace », « la jungle du Mexique / », « un tunnel noir \ »).

Les émotions suscitées par l'élément déclencheur ne sont pas souvent citées dans les productions. L'élément déclencheur correspond le plus souvent à un problème servant à produire les évènements subséquents (Godard, 2009).
Les différentes parties du récit sont enchevêtrées les unes dans les autres, particulièrement pour ce qui est des PI, et les actions se suivent sans être marquantes. Les sanctions/conséquences sont rares, ce qui est caractéristique des certaines de plusieurs récits narrés à chaine non focalisée (Schneuwly et Dolz, 1987) : des évènements sont présentés en fonction d'un problème, mais celui-ci n'est pas résolu dans le courant de la narration. L'état final est généralement lapidaire, compris en une seule phrase.
Une analyse de correspondances multiples a été menée afin de dévoiler les ensembles sous-jacents parmi les mesures performatives et narratives/structurelles. La dimension 1 (22,49% de variance expliquée), est construite par les variables dialogues, élément déclencheur, sanction, état final et parole du conteur (interaction conteur et auditoire). Cette dimension sous-jacente laisse entrevoir un lien privilégié entre les parties du schéma narratif et deux variables propres au « raconter » : l'insertion de dialogues et la posture de conteur•euse. Les ouvertures et clôtures au récit, qui représentent, avec l'absence d'actions, la dimension 2 (15,29% de variance expliquée), posent de façon limitée l'implication des paramètres physiques de l'action langagière (ex. « Il était une fois », « Fin »). L'instauration du lien entre conteur•euse et destinataire, ou la prise en compte de la situation de communication, n'est pas maitrisée. Ces nouvelles dimensions illustrent l'idée que, dans notre corpus, la présence d'une superstructure narrative cohérente (Schneuwly et Dolz, 1987) n'est pas assurée par la présence d'une clôture et d'une ouverture.

3.3. Analyses des mesures linguistico-discursives et performatives

Nous avons soumis nos données PI et PF à des analyses t de student pour mesures répétées. Les variables relatives à la richesse du vocabulaire testées se sont révélées non-significatives, ce qui suggère que les élèves n'ont pas progressé sur ce plan. Il nous a semblé plus intéressant de mener une analyse des productions sur le plan linguistique par la caractérisation de l'ensemble des productions au design expérimental prétest-posttest.
Dans le but de retrouver des patrons de fonctionnement dans les données, nous avons procédé au croisement des mesures linguistiques et des mesures structurelles ou narratives, en excluant les mesures linguistiques qui étaient peu représentées dans les données. Le calcul du khi-deux a été effectué à l'aide d'un tableau croisant de neuf mesures linguistiques et performatives et les cinq grandes dimensions du récit ($\chi2(32) = 236,44$, $p < 0,001$). Notre analyse repose sur les contributions au khi-deux pour chacune des cellules, qui permettent de voir les surreprésentations et les sous-représentations dans les données et

d'expliquer comment les deux variables peuvent être liées. Les contributions au khi-deux résultent en effet de la différence entre les valeurs observées pour chacune des variables et valeurs théoriques, permettant de rendre compte de certaines tendances.

Les organisateurs textuels ont des contributions au khi-deux importantes, ce qui signifie probablement qu'ils sont pertinents et discriminants en ce qui a trait au lien entre les deux variables : mesures structurelles ou narratives et mesures linguistico-discursives. Nous pouvons notamment observer une surreprésentation du nombre d'organisateurs textuels au sein des états initial et final et une sous-représentation de dialogues, de discours direct et d'éléments faisant partie de la chaîne anaphorique. Les organisateurs textuels « un jour », « donc », « et après » introduisent souvent l'élément déclencheur.

Souvent, dans les PI analysées, le repérage des parties du récit est compliqué par la succession d'actions, et rares sont les occurrences de résolutions de problème qui nouent le récit (voir aussi Schneuwly et Dolz, 1987). Cette suite d'actions est souvent ponctuée d'organisateurs textuels simples, notamment « et pis », comme en témoigne la production suivante :

> il y a un monsieur et une dame qui viennent de se marier \ **et pis** : tout contents : / ils rentrera à la maison \ **et pis** soudain : ils entenda la le téléphone qui sonna \ le pa :pa va répondre \ **et** c'était quelqu'un qui demandait s'il pouvait all' danser dans un public parce que le parce que la personne qui devait danser était malade \ **et pis** il dit il réponda oui \ **pis** il s'est trouvé une dan :se sur internet pendant un tr :ès tr :ès long moment / pis à un moment il trouva / tout content \ . **et pis** … y y retéléphona en disant que là c'était euh que :: que : il avait trouvé \ … **pis** il y va danser dans le public \ gigantesque \ avec beau :coup beau :coup de musique \ **et pis** : la musique était finie **donc** y rentra à la maison \ **et pis** : il va se regarder lui-même danser à la télé \ (P., PI, 5e primaire)

Les occurrences de pauses d'une longueur importante (plus d'une seconde) sont surreprésentées au début des productions, au sein de l'état initial, et sous-représentées dans les autres parties du récit. Toutefois, un examen qualitatif plus fin des productions laisse penser que certaines pauses plus longues (celles au-delà de 4 secondes) ont tendance à se produire assez tard dans le récit. De plus, les pauses semblent marquer le passage d'une phase du schéma narratif à l'autre, en particulier dans les productions courtes, dans lesquelles une phrase orale, ou une proposition, suffit à représenter une portion du schéma.

Les éléments linguistico-discursifs caractéristiques de la syntaxe de l'oral, comme le double marquage du sujet (ex. « son camion il s'est cassé », SM, PF, 5e) la répétition avec effet d'emphase, ou dispositif clivé en « c'est qui » (ex. « c'était tous ses amis : qui l'avaient suivi pour sa fête d'anni :ver :sai :re », AG, PI, 5e) contribuent au khi-deux de façon mineure. Ces éléments semblent ne pas être particulièrement pertinents pour définir les liens entre mesures structurelles et discursives.

4. Conclusion et mise en perspective

Le présent article nous permet d'ébaucher deux types de conclusions : l'une est théorique, et se rapporte aux critères de progression et d'évaluation de l'oral ; l'autre est méthodologique, et fait référence à la méthode adoptée afin de procéder à l'analyse des données textuelles dans la présente recherche.

Nous avons proposé des pistes théoriques permettant de mieux circonscrire ce à quoi nous sommes en droit de nous attendre de la part d'élèves du primaire en ce qui a trait aux récits oraux narrés monogérés. Nos analyses, dans la mouvance des méthodes mixtes, rendent aussi compte de patrons de fonctionnement ou dimensions sous-jacentes propres à l'oral dans les récits spontanés des élèves.

Sur les plans méthodologique et épistémologique, les analyses descriptives effectuées sur les portions structurelles, linguistiques et performatives des productions ont permis d'asseoir l'importance de nous intéresser qualitativement aux données. Les analyses statistiques utilisées (khi-deux, t de student, ACM) nous permettent de vérifier que la grille atteste de variations interindividuelles, signes de son efficacité pour l'enseignement, mais aussi de son utilité comme outil de recherche. Les analyses qualitatives ont révélé des formes de progression.

La méthode adoptée permet de mettre en valeur les avantages pour les sciences de l'éducation à délaisser les positions puristes, dichotomiques, du quantitatif-qualitatif, pour aller vers la recherche de la pertinence théorique et méthodologique. Les méthodes mixtes sont en outre susceptibles d'enrichir les hypothèses interprétatives issues des analyses préliminaires et de les vérifier empiriquement.

Bibliographie

Bronckart J.-P. (1997). *Activité langagière, textes et discours. Pour un interactionnisme socio-discursif*. Paris : Delachaux et Niestlé.

De Weck G. et Rosat M.-C. (2003). *Troubles dysphasiques. Comment raconter, relater, faire agir à l'âge préscolaire*. Paris : Masson.

Diez-Itza E., Martinez V., Perez V. et Fernandez-Urquiza M. (2017). Explicit Oral Narrative Intervention for Students with Williams Syndrome. *Front Psychol*, 8, 2337. doi: 10.3389/fpsyg.2017.02337

Dolz J. et Schneuwly B. (2017 [1998]). *Pour un enseignement de l'oral. Initiation aux genres formels à l'école*. Paris : ESF.

Drijbooms E., Groen M. A. et Verhoeven L. (2017, Jul). Children's use of evaluative devices in spokenetwritten narratives. *J Child Lang*, 44(4), 767-794. doi: 10.1017/S0305000916000234

Gagnon R. et Dolz J. (2016). Enseigner l'oral en classes hétérogènes : quelle ingénierie didactique ? *Les dossiers des sciences de l'éducation*, (36), 109-129. doi: 10.4000/dse.1405

Godard L. (2009). Le développement du discours narratif oral d'élèves du primaire. *Revue québécoise de linguistique*, 23(2), 73-100. doi: 10.7202/603093ar

Griffin T. M., Hemphill L., Camp L. et Wolf D. P. (2004). Oral discourse in the preschool

yearsetlater literacy skills. *First Language*, 24(2), 123–147.
Hassan R. (2011). Analyser des performances : un examen à partir de la didactique de l'oral. *Recherches en didactiques*, 1(11), 143-162.
Kern S. (1997). *Comment les enfants jonglent avec les contraintes communicationnelles, discursives et linguistiques dans la production d'une narration*. Université Lumière-Lyon II, Paris.
Kraif O. (2012). *AnaText*. Grenoble : Université Stendhal Grenoble 3. Récupéré de http://phraseotext.univ-grenoble-alpes.fr/anaText/
Nonnon É. (2016). 40 ans de discours sur l'enseignement de l'oral : la didactique face à ses questions. *Pratiques*, (169-170). doi: 10.4000/pratiques.3115
QSR International Pty Ltd. (2020). NVivo (*distribué en mars 2020*). https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.
Schneuwly B. et Dolz J. (1987). La planification langagière chez l'enfant. Eléments pour une théorie. *Revue suisse de psychologie*, 46, 55-64.
The jamovi project. (2021). *jamovi (Version 1.6)*. Récupéré de https://www.jamovi.org

Complémentarité du quantitatif et du qualitatif en ADT. L'immigration dans les discours du Front National

Camille Bouzereau

Université Côte d'Azur, CNRS, BCL –
camille.bouzereau@univ-cotedazur.fr

Abstract
This text proposes a linguistic analysis of the Front National's discourses by articulating quantitative and qualitative approaches in the line of logometric studies. In this framework, the interpretative path progresses from the identification of a salient entry that crosses the statistical notions of specificity and banality to the analysis of the semantic environment of this lexis and ends with the observation of a sub-corpus of attestation extracted from key sentences. The article thus focuses on the transition from statistical findings to the description of semantic correlates and unfolds a progressive methodological protocol that moves from measurement to meaning.

Keywords: textual statistics, discourse analysis, specificity, banality, co-occurrence, key phrase

Résumé
Ce texte propose une analyse linguistique des discours du Front National en articulant les approches quantitatives et qualitatives dans la lignée des études logométriques. Dans ce cadre, le parcours interprétatif progresse du repérage d'une entrée saillante, soit le terme « immigration » dont l'usage dans notre corpus permet de croiser les notions statistiques de spécificité et de banalité, vers l'analyse de l'environnement sémantique de cette lexie et se termine par l'observation d'un sous-corpus d'attestations extrait de phrases clés. L'article porte ainsi sur le passage des constats statistiques vers la description des corrélats sémantiques et déroule un protocole méthodologique progressif qui chemine de la mesure au sens.

Mots clés: statistique textuelle, analyse du discours, spécificité, banalité, cooccurrence, phrase-clé

1. Introduction

Cet article articule les approches quantitatives et qualitatives dans la visée de décrire linguistiquement les discours du Front National (FN) sur la période contemporaine définie en sciences politiques comme la période de normalisation,

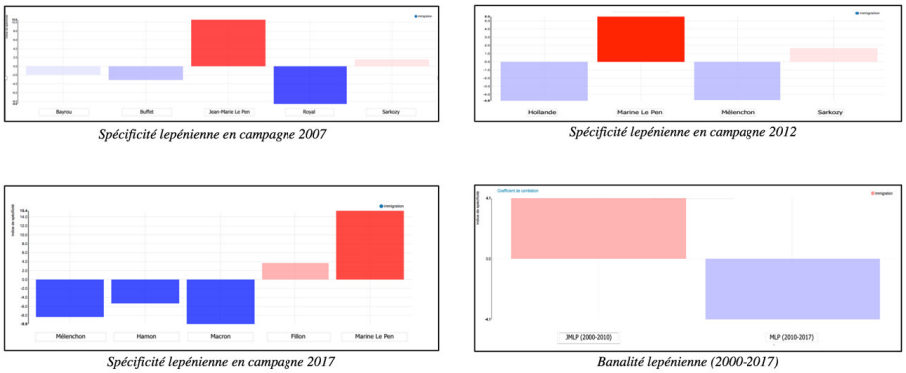
bornée des années 2000 à la fin des années 2010 (Dézé, 2012)[1]. Dans ce cadre, il propose un protocole méthodologique progressif croisant la linguistique de corpus assistée par ordinateur (Mayaffre, 2014) à l'appareillage conceptuel de la linguistique de l'énonciation et de la pragmatique (Jaubert *et al*., 2012). Ainsi, cet article part du postulat que la mesure a un sens et progresse, selon les perspectives ouvertes par F. Rastier (2011), de la constitution d'un corpus contrastif, du repérage de qualités saillantes pour aboutir à l'identification et à la caractérisation de formes qui se détachent du fond. Le traitement quantitatif est réalisé avec le logiciel Hyperbase (Brunet, 2011) et le corpus se divise en quatre bases de données dont la première propose un contraste interne au parti en comparant les discours de Jean-Marie Le Pen (JMLP) à ceux de Marine Le Pen (MLP) sur l'intervalle 2000-2017 (1 million de mots, 326 discours) ainsi qu'un contraste externe en comparant les discours lepéniens à ceux de leurs concurrents en campagne présidentielle de 2007, 2012 et de 2017 (2 millions de mots, 303 discours)[2]. Une fois le corpus constitué et les premières recherches effectuées se dégagent différents champs de pertinence dont le premier s'inscrit dans une approche microstructurelle attachée à l'analyse des spécificités lexicales. C'est dans cette approche que s'inscrit cet article qui suit un entonnoir heuristique en progressant du repérage de qualités saillantes situées à l'intersection de la spécificité et de la banalité (2) à l'analyse de l'environnement sémantique d'une lexie spécifique, soit le mot « immigration » (3) pour terminer par l'analyse d'un sous-corpus d'attestations extrait d'une fonction statistique d'Hyperbase (4).

2. L'*immigration*, au carrefour de la spécificité et de la banalité

Notre approche microstructurelle utilise la statistique occurrentielle et met en confrontation les notions de *spécificité* et de *banalité*. Si les spécificités identifient les mots caractéristiques d'un texte en comparant la fréquence de leur emploi dans un texte avec celle dans le reste du corpus, la banalité, depuis Tournier (1998 : 306), désigne les formes qui sont signalées mathématiquement sans spécificité : la banalité correspond aux formes dont l'indice de spécificité s'avère inférieur à +/-5%. Dans notre corpus, seul le terme « immigration » se situe à cette intersection : il s'agit d'un terme non seulement sur-utilisé par les

[1] L'article est issu d'une recherche doctorale qui a posé la question de l'existence d'une catégorisation linguistique pertinente pour caractériser les discours du FN (Bouzereau, 2020).
[2] Cette recherche alimente ainsi les études lexicométriques actuelles sur les discours FN qui s'attachent le plus souvent à des comparaisons intra-partisanes (Alduy 2015).

locuteurs lepéniens en campagne ((+10.6) chez JMLP en 2007, (+5.6) chez MLP en 2012, (+15.4) chez MLP en 2017) et d'un terme banal selon l'usage qu'en fait JMLP par rapport à celui de MLP ((+4.09) chez JMLP (2000-2010) et (-4.09) chez MLP (2010-2017)).



Spécificité lepénienne en campagne 2007

Spécificité lepénienne en campagne 2012

Spécificité lepénienne en campagne 2017

Banalité lepénienne (2000-2017)

Figure 1. L'immigration, un point de rencontre entre la spécificité et la banalité dans le vocabulaire lepénien

La lexie « immigration » représente ainsi un point d'entrée saillant pour comparer les usages que font les discours de JMLP face aux usages de JMLP et nous progressons, dans les sections suivantes, de l'analyse des différents environnements contextuels (3.) vers l'examen des supports discursifs les plus récurrents (4.).

3. Environnement sémantique d'immigration dans les discours du FN

Dans la lignée de Martinez (2003), la méthode cooccurrentielle articule les constats statistiques à la description de corrélats sémantiques. Cette statistique compositionnelle considère aujourd'hui que l'usage « prime sur le codage » et que « le sens n'est pas fait de référence (en dictionnaire) mais de différences (dans les corpus) » (Mayaffre 2014 : 16). Partant, la sémantisation du terme « immigration », dans notre corpus, s'appuie sur les données cooccurrentielles qui fonctionnent comme des points d'entrée saillants dans les parcours d'interprétation du sens et les profils cooccurrentiels étudiés révèlent un réseau sémantique commun pour les deux locuteurs lepéniens qui placent systématiquement l'immigration dans l'intervalle d'une relation causale. En témoignent les deux nuages de mots ci-dessous qui présentent les principaux substantifs lemmatisés, associés au terme « immigration » dans le discours lepénien sur l'intervalle 2000-2017.

Figure 2. Profili cooccurentiel d'immigration chez JMLP (2000-2010)

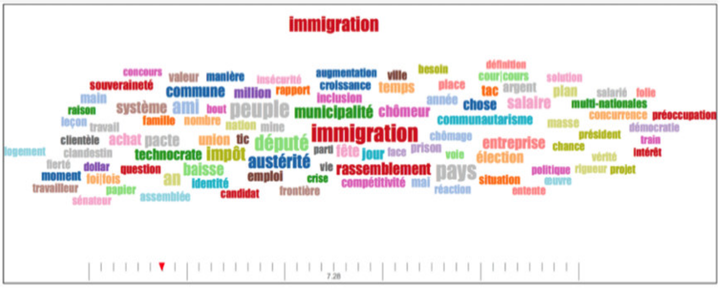


Figure 3. Profili cooccurrentiel d'immigration chez MLP (2010-2017)

L'immigration se situe ainsi entre la « cause » que représentent la « mondialisation », le « système », l'« union » européenne, les « politiques » en général et les « conséquence[s] » que sont « l'insécurité » pour JMLP, les problèmes d'« identité », le « chômage » et le « terrorisme ». Notons qu'aux lemmes « cause » / « conséquence », Marine Le Pen préfère la dichotomie « raison »/« solution ». Elle développe, en effet, un discours plus positif sur la thématique parce que démographiquement résoluble. Ainsi elle sur-utilise, par rapport à JMLP, les termes « compétitivité », « entreprise », « achat », « plan », « emploi », « valeur » et cela a une visée précise : faire du rejet de l'immigration un thème électoral. Dans l'environnement co-textuel d'immigration chez MLP, les termes de campagne tels que « municipalité » et « élection » vont effectivement se trouver surexploités par rapport à l'usage qu'en fait JMLP. S'il y a un profil commun aux deux locuteurs, certaines nuances sont donc observables. JMLP va effectivement plus loin dans la nomination : JMLP est le seul qui parle de manière significative de l'immigration de « peuplement », « de colonisation » et d'« invasion » quand Marine Le Pen intègre la thématique dans une stratégie électorale tout en systématisant un lexique anti-islamiste : l'environnement co-textuel d'immigration durant la campagne de 2017 révèle la sur-utilisation des lemmes « terrorisme », « fondamentalisme » et « communautarisme ». Notons enfin la présence de deux lemmes qui apparaissent dans les deux nuages de mots ci-dessus : dans le discours de Jean-Marie Le Pen (2000-2010), le lemme « nom » cooccurre avec « immigration », quand le

discours de Marine Le Pen (2010-2017) privilégie la cooccurrence avec sa « définition » :

> Nos débats, nécessaires, sur les problèmes institutionnels vous auront détournés quelques jours durant, des questions brûlantes qui se posent au pays et que vous connaissez bien [..]. **Elles ont des noms : chômage, immigration, insécurité, fiscalisme, dénatalité, santé, corruption morale et sociale**. JMLP, le 2 septembre 2001, à Aix-en-Provence.
> [...] quasiment l'intégralité ici des mouvements politiques français sont dans une vision post-nationale c'est-à-dire qu'ils considèrent que la nation est un concept dépassé et leur Europe a commencé comme un joli rêve et se termine comme un épouvantable cauchemar, elle se termine avec des frontières qu'on ne maîtrise plus, **une immigration qu'on ne maitrise plus, par définition**. MLP, le 28 avril 2014, à Douai.

En contexte, ces cooccurrents révèlent que le discours de JMLP systématise la nomination du mot dans le débat public (citation 1) quand celui de MLP systématise le parler FN sur le sujet. Une « immigration maîtrisée » serait, pour MLP, une immigration ayant lieu au sein de « frontières maîtrisées » : à partir du moment où l'Europe abdique ses frontières, notamment par la directive des travailleurs détachés, l'immigration ne pourrait plus être maîtrisée. Dans ce cadre, la conception européenne de l'immigration (maîtrisée) est contredite par les actes politiques et impose la conception lepénienne de l'immigration (non maîtrisée). Ce discours, tenu par MLP en 2014, a été préparé en amont, par JMLP. S'il ne dépasse pas le stade de la nomination, JMLP a imposé une certaine manière de parler de l'immigration dans le débat public. Les phrases évaluées, comme spécifiques de JMLP par rapport à MLP, permettent d'en rendre compte.

4. Les phrases clés, un sous-corpus d'attestations

Les phrases-clés désignent les passages caractéristiques d'un texte et leur calcul repose notamment sur les spécificités lexicales[3]. Leur examen distingue plusieurs récurrences dont la systématisation par JMLP, par rapport à MLP, des procédés de rectification et de renomination. Si ce résultat a un sens politique – JMLP systématise leur manière de parler de l'immigration pour que MLP en fasse un enjeu électoral – nous nous interrogeons plus précisément ici sur la corrélation entre la valeur statistique et la valeur pragmatique de ces phrases-clés. Dès lors, nous approfondirons l'examen de trois phrases clés extraites du discours prononcé le 13 janvier 2008 et qui mettent en relief ce procédé discursif. Ce

[3] Au moyen du modèle hypergéométrique converti en écarts réduits.

discours est circonscrit autour d'une critique contre la loi relative à la maîtrise de l'immigration, promulguée le 20 novembre 2007 – ce projet de loi avait été déposé au parlement sur l'initiative du gouvernement, dans la continuité de la campagne de Nicolas Sarkozy qui prônait durant sa campagne électorale le concept « d'immigration choisie ». C'est dans ce cadre qu'intervient la première phrase-clé :

> La loi Hortefeux, récemment adoptée, a réussi la prouesse d'accroître l'**immigration familiale, de faciliter l'immigration de travail, rebaptisée immigration choisie** et de maintenir **le flot de l'immigration illégale**. JMLP, le 13 janvier 2008 à Paris.

Dans cette citation, selon le locuteur, la loi serait en train d'accroître toutes les formes d'immigration. Dans ce cadre, JMLP dénonce le mensonge qui serait sous-jacent à la renomination sarkozyste : selon lui, « immigration choisie » remplacerait la dénomination « immigration de travail ». Il pointe, par le participe « rebaptisée » la prétendue création d'un nouveau concept, là où en fait, il s'agirait, d'un simple jeu d'étiquettes. Le cadre est ainsi posé pour amener *les bons mots*, jugés adéquats, selon JMLP et, c'est dans ce contexte que se manifeste la seconde phrase-clé :

> Au final, **il y aura plus d'immigration de travail, donc plus d'immigration familiale et plus d'immigration illégale**. C'est clair, l**a fameuse immigration choisie accroît l'immigration subie**. JMLP, le 13 janvier 2008 à Paris.

La dénomination sarkozyste « immigration choisie » est ici effacée au profit des syntagmes « immigration de travail », « immigration familiale » et « immigration illégale ». Puis, le SN « immigration choisie », parce qu'il est précédé de l'adjectif « fameuse » sonne bien comme n'appartenant pas au locuteur JMLP mais au discours d'un énonciateur autre (soit Nicolas Sarkozy). Et c'est le contraire avec l'expression « immigration subie » qui est quant à elle assumée par JMLP. Une fois les *bons mots* posés, JMLP peut amener ses propositions. Les discours du FN reposent systématiquement sur ce schéma binaire : avec une première partie systématiquement à charge et une seconde qui expose le discours FN comme le seul contre-discours. C'est dans ce cadre que JMLP actualise dans son discours, le néologisme discriminatoire *immigration-invasion* :

> La France qu'il aime, c'est la France brassée, par **l'immigration de masse**. La France qu'il aime, c'est la France américanisée. Dès lors, mesdames et messieurs, il ne peut qu'appeler de ses vœux un melting-pot à la française. Dès lors, **l'immigration-invasion va continuer**, comme d'ailleurs le déferlement de l'insécurité – l'une et l'autre sont liées. JMLP, le 13 janvier 2008 à Paris.

Jusqu'alors le discours se construisait contre celui des adversaires visant à disqualifier le contenu de leur proposition ainsi que leurs choix terminologiques. Dans cette citation il en va autrement : sans commentaire ou glose méta-énonciative et usant même de l'article de notoriété pour actualiser les syntagmes « immigration de masse » et « immigration-invasion » le discours FN apparaît comme une contre-doxa, épurée du discours des autres sur le sujet. En effet, ce néologisme « immigration-invasion » n'apparaît qu'en fin de discours, c'est-à-dire après avoir disqualifié les dénominations des autres et, surtout après avoir posé l'équation *immigration choisie = immigration familiale + immigration de travail + immigration illégale = une immigration de masse*. Les jeux d'étiquettes mis en relief par les phrases-clés sont donc représentatifs du discours de JMLP et occupent par ailleurs une place importante dans le déroulé de l'argumentation. Dans cette dernière phrase, le syntagme « immigration de masse », pris en charge par JMLP, désigne la conséquence de ce que Nicolas Sarkozy appellerait par euphémisme « immigration choisie » *a contrario* de JMLP qui parlerait sans se voiler la face d'« immigration de travail ». Ainsi, à cette nomination présentée comme euphémistique, le discours FN souligne qu'il adopte une nomination qui *collerait au réel*, à savoir celle d'« immigration de travail » et en l'occurrence, cette immigration serait la porte ouverte à l'*immigration-invasion*. La création du néologisme fait sens hors contexte et sa motivation sémantico-référentielle se justifie en ce que cette alliance lexicale porte en elle la réception polémique qu'elle va engendrer, soit le débat polarisé opposant ceux qui assertent qu'il faut bien dire les choses comme elles sont, non comme on nous les présenterait et ceux qui refusent de telles images. L'emploi de notoriété vise ainsi à présenter la conception lepénienne comme étant la seule qui vaille et, dans cette perspective, il n'y aurait plus qu'une seule immigration : celle qui est massive et par définition une invasion.

L'acte de nommer constitue un support discursif important : ces procédés permettent à JMLP de mettre en exergue certaines dénominations pour les contredire, s'y opposer et proposer d'autres manières de nommer un concept de la vie politique. Si le mot « immigration » est surinvesti dans les discours FN, son environnement lexical construit une thématique précise. En discours, le but des candidats FN n'est pas seulement de charger négativement le concept mais bien d'imposer le rejet de l'immigration, comme la seule manière d'en parler.

5. Conclusion

L'observation des spécificités lexicales met en évidence que le terme « immigration » occupe une place fondamentale dans le vocabulaire du Front National d'une part et, dans la construction discursive de l'idéologie d'extrême droite, d'autre part. Ensuite et, partant de phrases-clés construites autour du terme « immigration », nous avons montré que ces énoncés calculés clés statistiquement

sont aussi souvent clés pragmatiquement, dans l'organisation globale du discours. Le retour au texte montre que le terme est toujours chargé négativement et apporte ensuite de nouvelles dénominations et donc de nouvelles représentations. Ces représentations participent à façonner un discours et servent une fin politique. Depuis plus de 20 ans, le FN essaye effectivement d'imposer le rejet de l'immigration comme étant la seule manière d'investir le thème. Notre article est ainsi passé du repérage d'une qualité saillante à l'analyse de corrélats sémantiques et a bien nécessité ce faire de la complémentarité des approches quantitative et qualitative.

Bibliographie

Alduy C. and Wahnich S. (2015). *Marine Le Pen prise aux mots*, Paris, Seuil.

Bouzereau C. (2020). Thèse de doctorat soutenue à l'Université de Nice Côte d'Azur le 27 novembre 2020. *Doxa et contre-doxa dans la construction du territoire discursif du Front National (2000-2017)*.

Brunet É. (2011). HYPERBASE Manuel de référence. 9.0 Distribué par *l'UMR 7039*. En ligne. URL : http://ancilla.unice.fr/bases/manuel.pdf (consulté le 10 mars 2022).

Dézé A. (2012). *Le Front National à la conquête du pouvoir.* Paris Armand Colin.

Jaubert A., Lopez J., Marnette S., Rosier L. and Stolz C. (2012). Citations I, Citer à travers les formes. Intersémiotique de la citation. *Academia L'Harmattan*.

Martinez W. (2003). *Thèse de doctorat soutenue à Paris 3*. Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels.

Mayaffre D. (2014). Plaidoyer en faveur de l'analyse de données co(n)textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014). *Journées internationales d'Analyse statistique des Données Textuelle*, Paris Inalco-Sorbonne nouvelle, p. 15-32.

Rastier F. (2011). *La mesure et le grain*. Paris Champion.

Tournier M. (1998). Banalité spécifique dans le discours syndical : adverbes et adjectifs de l'insistance. *Mots chiffrés et déchiffrés : mélanges offerts à Étienne Brunet*, Paris Champion, p. 305-316.

Semantic and morphological properties of gender assignment in Italian: distributional data on common nouns

Giulia Bracco, Maria De Martino, Alessandro Laudanna

University of Salerno – gcbracco@unisa.it mdemartino@unisa.it alaudanna@unisa.it

Abstract
The paper describes the interplay between semantic and morphological properties of grammatical gender and the distribution of Referential (semantically based) Gender vs. Grammatical Gender in contemporary written Italian.
Quantitative data on common nouns were obtained by exploiting and proving further annotation to the DeGNI lexical database (Declinazione e Genere dei Nomi Italiani, De Martino *et al*., 2018; 2019) in order to isolate naturally gendered nouns such as human and animal agents and describe morphological and lexical phenomena of "gender shifting".

Keywords: grammatical gender, referential gender, distributional data.

1. Semantic and morphological properties related to grammatical gender

There is general evidence for the influence of grammatical gender in the cognitive processing of words (Vigliocco *et al*., 2005) and in the mental representations of the referents (Vygotsky, 1986). Studies across languages show that grammatical gender is not expected to have semantic connotation per se: words with the same referent show different genders through languages: "the sun" is masculine in Italian (*il sole*), feminine in German (*die sonne*) and neutral in Polish (*to słońce*). A difference is made between systems of lexical-semantic origin, and gender systems which arise for morphological or syntactic reasons (Fodor, 1959). Differences in semantic systems (referred to as "semantic gender") reflect different correlations with the grammatical systems.
While in Genderless Languages (e.g., Finnish) grammatical gender is not represented, in Natural Gendered Languages (e.g., English, Scandinavian), grammatical gender has semantic foundations and nouns are gendered only in case of sexed referents, mainly through pronouns (such as "he" or "she" in English) and nouns have no grammatical marking of gender.

Conversely, in Grammatical Gendered Languages (e.g. Italian, German) grammatical gender features (feminine, masculine and sometimes neuter) are also attributed to inanimate objects. In this case, the grammatical gender is not referential and is mostly arbitrary. However, a tendency for inanimate objects and/or concepts to be mainly assigned to a given grammatical gender can be described. For instance, in Italian, fruits are mainly feminine (*mela*, apple), while fruit trees (*melo*, apple tree), tend to be masculine (Thornton, 2001). Languages with a grammatical gender system have also a semantic gender system: when a noun refers to human beings or animals, it generally reflects the gender/sex of the individual in question. In contrast, in Grammatical Gendered Languages grammatical gender is an inherent property of nouns with mainly syntactic and morphological functions (Chomsky, 1965) and it affects case markings, articles and all the other forms of agreement provided by the language. It has been pointed out that different levels of information associated with gender in different languages could possibly affect cognitive processing of words.

According to some theories, semantic features of Human/Animal vs. Inanimate are the most relevant in conceptualization, and semantic dichotomies such as Animate vs. Inanimate, Human vs. Non-Human, Masculine vs. Feminine are crucial in the natural classification of nouns (Dahl, 2000). Since the morphological features have deep implications in processing semantic information, grammatical gender marking is expected to carry connotative meanings of femininity and masculinity (Vygotsky, 1986). By the same token, it has been shown that grammatical gender can affect the processing and the categorization of inanimate objects (Cubelli *et al*., 2011), also on semantic bases (Chini, 1995).

The Sex and Gender Hypothesis (Vigliocco *et al*., 2005) maintains that in languages with two grammatical genders, such as Italian, grammatical gender may affect also nouns for which semantic gender is not relevant, such as inanimate objects[1]. According to Lepschy (1989 one hypothesis is that in Italian, a language with two genders, this feature has mainly lost its semantic/referential function and it does not involve that we could, for instance, animistically, refer a biological essence to a particular entity through the grammatical gender of its noun. Actually, in Grammatical Gendered Languages such as Italian, grammatical gender is necessary for the agreement of the totality of nouns, adjective, participles or articles, whereas nouns referred to sexed entities are a small semantic core. By this token, it would be possible to hypothesize that, in

[1] In Italian, no neutral gender is provided: living entities can be semantically affected by grammatical gender and inanimate objects can be either masculine (e.g., *libro*, book) or feminine (e.g., *penna*, pen)

the grammatical gender system of Italian, the semantic range is limited, even though, at least to our knowledge, no distributional data are available to circumscribe the phenomenon.

On the other hand, it has been demonstrated that the distributional association between grammatical gender and other gender-related morphological properties of nouns affects lexical access (De Martino *et al*., 2011; 2017; Paolieri *et al*., 2011; Russo *et al*., 2021); it is worth to remind here, that in order to avoid possible confounding with semantic processing, a set of psycho-(neuro)linguistic data exploring the role of grammatical gender mostly focuses on nouns referring to objects, concepts or asexed entities. Meanwhile, there is an increasing sensitivity to ensure that lexicon equally encodes natural gender of humans, and to avoid sexist biases in language (see Sabatini, 1987). It has been shown that language can reflect the processing of gender stereotypes (Wood and Eagly, 2002; Cacciari and Padovani, 2007; Cacciari *et al*., 2011): marking the animate noun for the opposite sex is the main linguistic approach to avoid or reduce inequality in individual gender representation.

The gender shifting phenomenon can be defined as an either inflectional or derivational process that allows to change the referential gender. It is a highly productive process in synchrony but nevertheless, it can be difficultly detected since lexical databases often lemmatize words in the citation form; for instance, the masculine noun form in Italian (Thornton, 2003). In Italian, the gender shift is formally realized at a morpho-phonological level, where it is not possible to disentangle the information of grammatical gender of a noun from its declension class and its meaning (Thornton, 2003). A change in the final vowel (e.g., *nonn-o/nonn-a*, grandfather/grandmother) or the apposition of a derivational affix (*attore/ attrice*, actor/actress) allow to obtain the respectively gendered referent. In these cases, the noun preserves the root. In other cases, the base form changes, and the gender shift is realized at a lexical level: two independent gender nouns represent a complementary relation between two different sexed identities (*fratello/sorella*, brother/sister, *moglie/marito*, wife/husband).

If has been hypothesized that semantic gender does affect nouns referring to inanimate referents; conversely, nouns referring to animate entities always undergo grammatical gender system. In the following paragraphs, a description of the gender shifting phenomena as present in the Italian gender system is provided and the distribution of such phenomena across nouns with grammatical vs. biological gender is analyzed. In our view, distributional properties of gender marking could be useful to describe semantic gender system in Italian.

2. Natural and Grammatical Gender Shifting in Italian

According to grammars, (Dardano and Trifone, 1995; Serianni, 2004) *mobile nouns* modify their natural gender simply changing the final vowel: gender

implies phenomena of inflection consistently with morpho-phonological rules in word formation. The root is mainly preserved.

In Italian the citation form of nouns is mostly the masculine singular, and the class of nouns which shifts gender by changing the final vowel is highly productive.

- Masculine nouns ending in *–o* mainly change the final vowel in *–a* (*zio* (M), *zia* (F), *aunt*, common gender in English, *deputato* (M)/*deputata* (F), deputy, common gender in English);
- masculine nouns ending in *–e* shift to feminine by taking the final vowel *–a* (*cameriere* (M)/*cameriera* (F), waiter/waitress).

Gender shifting can also involve phenomena of derivation through affixation:

- a class of masculine nouns ending in *–e* shifts to feminine by taking the suffix *– essa* (*conte/contessa*, earl/countess), or *–ina* (*eroe/ eroina*, hero/heroine);
- some masculine nouns ending in *–o* take the suffix *– essa* (*soldato* (M), *soldatessa* (F), soldier, common gender in English);
- masculine nouns ending in *–a* take the affix *–essa* (*poeta* (M)/ *poetessa* (F), poet, common gender in English);
- masculine nouns ending in *–tore*, take the affix *–trice* (*attore* (M)/ *attrice* (F), actor/actress), while masculine nouns in -sore shift to feminine through the affixes - *essa* (*professore* (M)/ *professoressa* (F), professor, common gender in English), or -*itrice* (*traditore* (M)/ *traditrice* (F), traitor, common gender in English).

In some cases, gender shifting leads a noun to change its declensional class:

- in common gender nouns in *–a* and in *–ista* the declensional class is either *a/e* or *a/i* according to the natural gender (*artista*/ *artiste* F, *artista*/ *artisti* M, artist/artists, common gender in English). Note that these words (*atleta*, athlete, *artista*, artist) also show their natural gender in the context (*la brava atleta* (F), *il bravo atleta* (M), the good athlete, common gender in English).

Gender shifting can also be realized through a different lexical entry, an independently gendered noun. These pairs are also defined “complementary”. Different kind of relationships between Grammatical Gender (GG) and Natural Gender (NG) can be described:

- in Independent Gender Nouns (*padre/madre*, father/mother, FGG; *moglie /marito*, wife/ husband, *mucca/toro*, cow/bull), Grammatical Gender matches up with Natural Gender;
- in Common Gender Nouns, Natural Gender can often be detected in the context (*la nipote* (F) *il nipote* (M), niece/nephew). In the same class we can find nouns derived from past participles (*la cantante* (F)/ *il cantante* (M) the singer);
- nouns for animals (e.g., volpe, fox) often refer to both male and female

exemplars and semantic markings can be completely absent; conversely, the information of grammatical gender is not semantically based but it has always a crucial role in the agreement (*la volpe affamata* FGG, the hungry fox);
- in a small class of nouns there is a mismatch between GG and NG: *soprano* refers to female singer (FNG) but it is a grammatically masculine noun (MGG).

We exploited DeGNI (De Martino *et al*., 2019; 2018), a lexical database encoding 23.619 nouns taken from the CoLFIS (Corpus e Lessico di Frequenza dell'Italiano Scritto, Bertinetto *et al*., 2005) database, to extract distributional data and cross multiple parallel information on the main features of Italian nouns such as grammatical gender, number and declensional class. Furthermore, additional annotations for DeGNI records were provided in order to isolate naturally gendered nouns such as human and animal agents. The aim was to describe the interplay between grammatical gender and natural (referential) gender, by closing off morphological phenomena such as inflection, or "gender derivation by affixation", and lexical phenomena such as "independent gender nouns".

3. Distributional data on animate common nouns in Italian

Following Dahl (2000), we distinguished between grammatical gender and referential gender by contrasting nouns referring to Inanimate Referents (objects or concepts) with nouns referring to Human/Animal agents. This latter class represent 21% of the 23.619 nouns from the DEGNI Database (95% Human Agents, 5% Animals). As far as the cluster Human/animal agents is considered, 50% of them are masculine (Table 1).

Table 1. Referential Gender of Human/Animal Agents

Common Gender Human/animal agents (e.g., *volpe*, fox, *adolescente*, adolescent)	35%
Feminine Gender Human/Animal Agents (*modella*, model, FRG)	15%
Masculine Gender Human/Animal Agents (*operaio*, worker, MRG)	50%

Table 2. Final Vowel associated with Common Gender Human/Animal Agents.

Nouns ending in *-a* (*atleta,* athlete)	36%
Nouns ending in *-e* (*fuoriclasse,* champion)	44%
Nouns ending in *-o* (*senzatetto,* homeless)	7%
Other endings (*bebè,* baby)	13%

It has been argued that (masculine or feminine) lexical representation of an individual reflects change in the thoughts and attitude about gender stereotypes (see Sabatini, 1987). Furthermore, note that in Italian the masculine singular is

often the citation form, and this might affect the lemmatization of corpora (Thornton, 2003). Interestingly, within the group of common gendered nouns, ending in *–o* is the smallest cluster (Table 2), even though it is ortho-phonologically related to the (masculine, singular) citation form.
Among the group of Human/animal Agents, we can observe that in the majority of nouns (65%) Referential Gender matches Grammatical Gender (*gemello* (MRG, MGG)/*gemella* (FRG, FGG), twin, common gender in English). Only a few nouns show a mismatch between GG and RG (*soprano* FRG, MGG), while 35% of Human/Animal Agents are Common Gendered nouns (*volpe*, fox, *adolesc*ente, adolescent, Table 3).

Table 3. Human/Animal Agents nouns showing a match/mismatch between Referential Gender and Grammatical Gender.

Common gender nouns	35%
Referential Gender = Grammatical Gender	65%
Referential Gender ≠ Grammatical Gender	0%

Table 4 shows the distribution of the different phenomena involved in gender shifting[2]. Only 5% of nouns have a new lexical entry for the opposite gender. Inflection (change in the final vowel) appears to be the most extensive morphological phenomenon (71%).

Table 4. Distribution of inflectional, derivational and lexical phenomena in gender shifting.

Inflection (*nonna*, grandmother)	71%
Derivation/affixation (*duchessa*, duchess)	24%
Lexicon (*sorella*, sister)	5%

4. Conclusions

If we assume the thesis that in Grammatical Gendered Languages gender has a semantic function only for animate beings (Lepschy, 1989) such as humans or animals, Referential Gender might be considered a minor phenomenon in Italian, whereas Grammatical Gender has morphological and syntactic functions for the totality of nouns (animate and inanimate referents). Distributional data show that

[2] Note that a small cluster (4%) of Human/Animal Agents are nouns that are semantically referred to one (masculine or feminine) individual and do not have a referent belonging to the opposite gender (*ninfa*, nymph, F, *evangelista*, evangelist, M).

gender shifting is mainly associated with morphological (particularly inflectional) phenomena, and affects declensional class information for both the genders in Italian. Quantitative data show that in Italian gender is mainly encoded as a morphological and grammatical feature.
To our view, distributional properties of grammatical gender should be taken into account to describe phenomena of prevalence or gender bias in contemporary written Italian.

References

Bertinetto P.M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., and Thornton A.M. (2005). Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS). http://alphalinguistica.sns.it/CoLFIS/CoLFIS_Presentazione.htm

Cacciari C. and Padovani R. (2007). Further evidence of gender stereotype priming in language: Semantic facilitation and inhibition in Italian role nouns. *Applied Psycholinguistics*, 28: 277–293.

Cacciari C., Corradini P., Padovani R. and Carreiras M. (2011). Pronoun resolution in Italian: The role of grammatical gender and context. *Journal of Cognitive Psychology*, 23: 416–434.

Chini M. (1995). Genere Grammaticale e Acquisizione. Aspetti della Morfologia Nominale dell'Italiano l2. Franco Angeli.

Chomsky N. (1965). *Aspects of the Theory of Syntax*. M.I.T Press.

Cubelli R., Paolieri D., Lotto L. and Job R. (2011). The effect of grammatical gender on object categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37: 449-60.

Dahl Ö. (2000). Animacy and the notion of semantic gender. In Unterbeck, B., Rissanen, M., Nevalainen, T, and Saari, M. editors. *Gender in Grammar and Cognition*. Mouton de Gruyter.

Dardano M. and Trifone P. (1995). *Grammatica Italiana. Con Nozioni Di Linguistica*. Zanichelli.

De Martino M., Bracco G., Postiglione F. and Laudanna A. (2018). *Declinazione e Genere dei Nomi Italiani: DeGNI, una banca dati lessicale*. http://www.lapsus.unisa.it/ricerca/index

De Martino M., Postiglione F., Bracco G., and Laudanna A. (2019). Declinazione e genere dei nomi italiani: Degni, una banca dati lessicale. *Giornale italiano di psicologia*, 46: 247-268.

De Martino M., Bracco G. and Laudanna A. (2011). The activation of grammatical gender information in processing Italian nouns. *Language and Cognitive Processes*, 26: 745-776.

Fodor I. (1959). The origin of grammatical gender. *Lingua*, 8: 186–214.

Lepschy G. (1989). *Nuovi saggi di linguistica italiana*. Il Mulino.

Pescuma V. N., Zanini C., Crepaldi D. and Franzon F. (2021). Form and function: a study on the distribution of the inflectional endings in Italian nouns and adjectives. *Frontiers in Psychology*, 12: 4422.

Sabatini A. (1987). *Raccomandazioni per un uso non sessista della lingua italiana*. http://www.funzionepubblica.gov.it/media/277361/linguaggio_non_sessist

Serianni M. (2004). *Grammatica italiana*. UTET.
Russo A. G., Esposito F., Laudanna A., Mancuso A., Di Salle F., Elia A. and De Martino M. (2021). The neural substrate of noun morphological inflection: A rapid event-related fmri study in Italian. *Neuropsychologia*, 151: 107699.
Thornton A. M. (2001). Some reflections on gender and inflectional class assignment in Italian. In Schaner-Wolles C., Rennison, J.R. and Neubarth, F. editors, *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*. Rosenberg & Sellier.
Thornton A. M. (2003). "L'assegnazione del genere in italiano", in Sánchez Miret, F. editor, *Actas del XXIII Congreso Internacional de Lingüística y Filología Románica, a cura di, vol. I, Tübingen, Niemeyer*, pp. 467-481.
Vigliocco G., Vinson D.P., Paganelli F. and Dworzynsk K. (2005). Grammatical gender effects on cognition: implications for language learning and language use. *Journal of Experimental Psychology*, 134:501-20.
Vygotsky L.S. (1986). *Thought and Language*. M.I.T. Press.
Wood W. and Eagly A. H. (2002). A cross-cultural analysis of the behavior of women and men: Implications for the origins of sex differences. *Psychological Bulletin*, 128: 699-727.

The development of word frequency distribution in first language acquisition. An analysis on a spoken language corpus of French children

Andrea Briglia[1], Massimo Mucciardi[2], Giovanni Pirrotta[2]

[1]Sorbonne Université – andrea.briglia@sorbonne-universite.fr
[2]University of Messina – massimo.mucciardi@unime.it
giovanni.pirrotta@unime.it

Abstract
In this paper we present a study on the word-frequency distribution development in French speaking children, which aims to evaluate how their lexical output is related to a standard word-frequency distribution: Zipf's law. We adopted a set of spoken language transcripts of French children named CoLaJE: by using Python tools we turned original transcripts into strings that allowed us to estimate the exponential parameter of the word-frequency distribution (*alpha*) for each child, as well as for parental input. We show how *alpha* values tend to converge to 1 during later development which is coherent with current literature. We also estimate the exponential parameter for parental input and we found that Spearman's rho shows a fairly positive correlation between child's *alpha* and parents' *alpha* in later ages. Finally, we discuss our results in the light of previous studies on the CoLaJE corpus and we compare the obtained values to similar works on children's spoken language transcripts that were sampled in an analogous way, before outlining possible future directions.

Keywords: Zipf's Law; Spearman's rho; Language Acquisition, Natural Language Processing

1. Introduction

Zipf's law has been at the center of a long-lasting scientific debate since its discovery at the beginning of the last century by French stenographer Estoup (Zipf, 1949). This rank-frequency distribution is an inverse relation found in an approximately similar way in a number of human and natural phenomena (Ferrer i Cancho and Solé, 2001). In mathematical terms, this means that the rank of a word is approximately inversely proportional to its frequency and so produces a power law probability distribution. As far as we know, there is not a statistically

significant difference in the shape of the plot between written and oral language, as it was suggested by a study on short oral samples similar to those used for our study (Ridley and Gonzales, 1994). Recent advances in NLP have increased understanding of this relation (Piantadosi, 2014) by pointing out how the universal word distribution would be an equilibrium between communication effectiveness and cognitive load that is unconsciously reached by listeners and speakers in a collaborative way (Lestrade, 2017). To our knowledge, research is still lacking regarding the emergence of a Zipfian word frequency distribution in French language during childhood. Our hypothesis is that some sort of pattern, tendency or regularity - which is to say, something that could be interpreted as a preliminary form of a Zipfian distribution - should be found in child language during the acquisition process. Preliminary results on these subjects suggest that a Zipf-like statistical distribution progressively emerges from around 30 months of age. In an analogous way to adults, stop words (especially determiners) are more frequent than nouns and verbs. However, if we look at the transcriptions, there are a significant number of exceptions that we claim to be determined by context-specific situations (such as "maman" and "papa"), as well as the frequent naming of a toy that was constrained to a specific situation (it is important to bear in mind that our study is based on one-hour samples of spoken language). When we compare these results to previous studies on these children (Morgenstern and Parisse, 2012; Mucciardi *et al*., 2021), it could be said that children acquiring their native language in a faster way (e.g Madeleine) showed an earlier Zipfian-shaped distribution (i.e., closer to alpha = 1) compared to the other children.

2. Data and NLP tools

CoLaJE (Morgenstern and Parisse, 2012; ver. 2.4) is a database made up of eight children[1] recorded *in vivo* one hour per month from their first year until they were approximately five. For each child there are nearly 8000 sentences and 20000 words with a mean length of utterance of three words on average. Child-directed speech has also been recorded and it is transcribed by using FAT and MOT lines. As adult input is of primary importance (Goodman, 2008) we decide to calculate the *alpha* for the parents too. This means that we calculate the sum of every occurrence of FAT (father) and MOT (mother). We have not, however, considered occurrences of OBS (the researcher filming the conversation), because this person spends just one hour per month with the observed child, so her language does not influence child development in a significant way. Adults in CoLaJE corpus are

[1] We decided to include the child named Philippine, even though her first recording is at 4 years old.

quite often speaking in the so-called "motherese": it is difficult to estimate how this influences the emerging word-frequency distribution in the examined children. This set of spoken language transcripts is considered statistically representative for studies in language acquisition (Yamaguchi, 2018). We analyze the corpus by using the following libraries: PyLangAcq (Jackson *et al*., 2016) for the manipulation of CHAT format (Mac Whinney, 2000) and Stanza (Qi *et al*., 2020) for the Part Of Speech (POS) tagging of children spoken utterances. Data are available in three different types[2]: CHI represents what the child should have said according to the adult norm in standard French orthography; MOD represents what the child should have said according to the adult norm in International Phonetic Alphabet and PHO represents what the child actually said in IPA. To obtain our results, we decided to choose CHI for a number of reasons. First, it is the only transcription available for all children, whilst MOD and PHO are available only for Adrien and Madeleine. Secondly, PHO poses more substantial problems for the purposes of the study, despite its advantage as the only transcription that exactly corresponds to real language. Indeed, a given word can be pronounced in many different ways with different degrees of variations, which makes computations complex: it is difficult to establish when a word means what it was meant to mean for a child, and to what extent different variated forms refer to the same entity, especially in earlier ages (Vihman, 1994). We thought that this choice would have biased our word-frequency counts, resulting in too many different word types that in most cases do not really represent different named entities. On the other hand, CHI represents what the child says in its correct form, which gives us an adult-biased interpretation of child vocabulary. It has nonetheless the advantage of gathering words that differed on what was imagined to be a phonetic level in a stable and fairly reliable way that in the majority of cases would not affect the lexical level. By proceeding in this way, we accepted a heavy reliance on the transcribers' choices.

3. Results and discussion

For every child, we estimated the *alpha* parameter according to the well-known

$$\mathrm{f}(r) \sim \frac{1}{r^{\alpha}} \tag{1}$$

Zipf's law (Zipf, 1949; Mandelbrot, 1961) showed in formula (1):

[2] Here is an example:
https://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data4/colaje/adrien/ADRIEN-21-3_00_15/ADRIEN-21-3_00_15.tei_corpo.xml

where r is the ranking of the word (from common to rare), f(r) is its frequency distribution for words in the child corpus and α (*alpha*) is the exponent of the law (a parameter defining the shape). We analyze a total of 237 corpora covering a developmental temporal range from 9.23 to 83.90 months. Figure 1 shows one of the 237 estimated models[3]. For each corpus examined we estimate the *alpha* parameter, obtaining the results shown in Table 1 and Figure 2 where alpha values are fitted according to a smoothing spline (Fan and Gijbels, 1996)[4]. If we look at figure 2, we can observe how the variability around alpha=1 is higher in earlier ages. It is possible to observe how the *alpha* parameter converges towards the value of 1 over time. In particular, after 48 months (data not shown) the *alpha* average is almost 1 (*i.e. 0.99*). We remind the reader that when *alpha* goes above 1, this means that there is a decrease in lexical richness, and vice versa: while *alpha* goes below 1 there is a growing lexical richness. Despite a limited vocabulary and an undeveloped syntax, we observe that the cognitive underpinnings behind the universality of Zipf are at play from the very beginning. The progressive emergence of a Zipf-like probability distribution in children can be influenced by an increased exposure to their parents' input. In our results, this appears by estimating Spearman's non-parametric correlations (Table 2). In fact, during the initial age of the children (<= 24 months), the correlation between the *alpha* parameter of the children and the *alpha* parameter of the parents is almost zero (0.005). With increasing age, the correlation appears to be significant and increases between 24 and 48 months (0.221). From the age of 48 months, it grows to a significant value of 0.330. We find these results coherent to previous findings on the same corpus (Mucciardi, 2021), in which we observe[5] two graphs representing the evolution of POS tags. We observe that for both children (despite the fact that the girl learns faster than the boy) there is a clearly significant increase in the quantity and quality of speech production from two-and-a-half-year-old (30th month): this is coherent with what we observe in Figure 2, as well as with the difference in the Spearman's values for the same age range over time. It is coherent in the sense that three different ways of measuring child language development (*i.e*. alpha evolution over time, Spearman correlation and the evolution of POS tags) show a similar convergence in which we can see how

[3] All the 237 models estimated are highly performing in terms of goodness of fit and are available upon request. Data analysis was performed using STATA ver. 15.0 software.

[4] We use a spline function with a kernel function= Epanechnikov and bandwitdh=5 (see for more details Fan and Gijbels, 1996).

[5] http://advanse.lirmm.fr/EMClustering Adrien (dataset 1) and Madeleine (dataset 2). Click on the bottom-left part to choose to visualize the "evolution of POS tags" in absolute values (default is in relative values).

children's utterances are becoming closer to adult language both in quantitative and qualitative terms. Furthermore, these results are in line with the graphs produced by the authors who led the constitution of the CoLaJE corpus: as we can see in their seminal paper (Morgenstern and Parisse, 2012), the mean length of utterance, type/token ratio and the number of utterances per hour evolve in a similar fashion. The progressive emergence of different grammatical categories (Mucciardi, 2021) is coherent to the graph proposed in the review article cited in the introduction (Piantadosi, 2014, fig.7). Additionally, analogous gaps between children with different paces of development can be observed in both

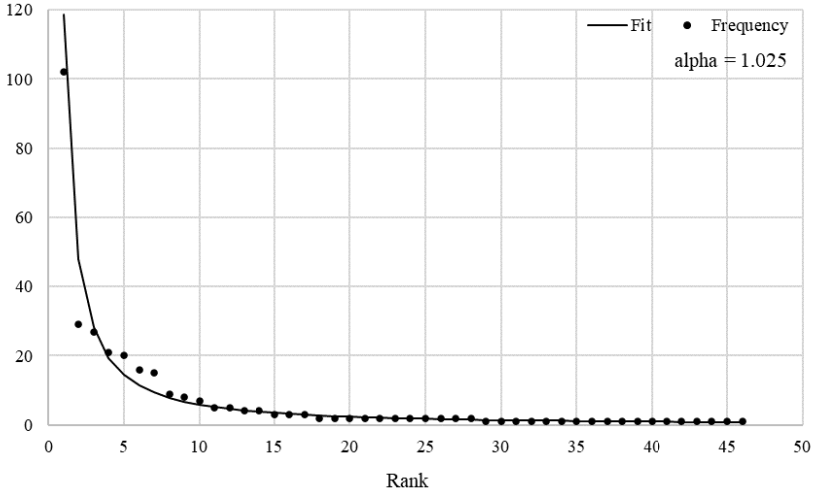


Figure 1. Estimated model for Theophile (age = 27 months): observed (frequency) and estimated (fit) values – R2adj=0.97 Anova test F=1416 (p<0.001)

graphs, as is the case for Madeleine and Adrien, while at around five-years old, children usually display less variability between their outputs (*alpha* values, mean length of utterance or type/token ratio).

Table 1. Descriptive statistics of alpha values on 8 children analyzed

Child	Mean (*alpha*)	N= corpus	Std. Deviation	Age (range in
ADRIEN	1.194	27	0.199	15.30 –
ANAE	1.021	32	0.082	16.67 –
ANTOINE	0.965	43	0.168	9.23 – 75.27
JULIE	1.010	27	0.179	11.60 –
LEONARD	1.024	14	0.080	20.30 –
MADELEINE	1.062	34	0.117	12.17 –
PHILIPPINE	0.907	20	0.190	56.90 –
THEOPHILE	1.035	40	0.181	14.33 –
Total	1.028	237	0.172	9.23 – 83.90

Table 2. Spearman correlations between children alpha and parental alpha by age group

Spearman's rho correlation		
Age <=24 (months)	Age >24 and <=48	Age >48 (months)
0.005	0.221*	0.330*
**p<0.05*		

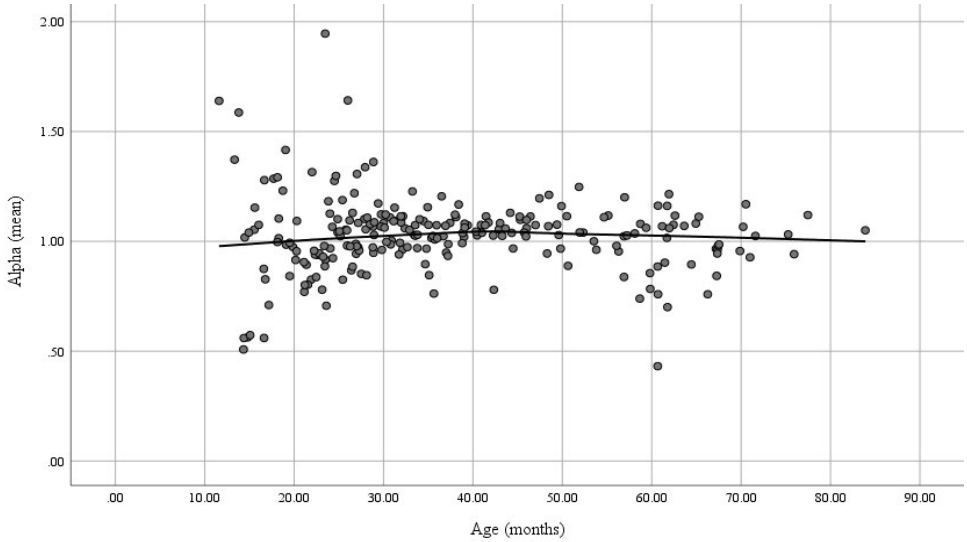


Figure 2. Temporal trend of the alpha parameter for the 8 children

4. Conclusion and future directions

To explore exponent evolution in child language development, we adopt as a reference a previous work (Baixeries, 2013) in which the authors challenged the assumption according to which the exponent of Zipf's law remains constant independently from the complexity of the communication system. They do so by pointing out how in children speaking Germanic languages the exponent "evolve from a high value of alpha to the value of alpha of adults at least from about 20 months onwards" (Baixeries, 2013). As for our study, these results have been tested on CHILDES data (Mac Whinney, 2000), which are similarly sampled spoken language transcripts. The results we obtained on French are on the one hand similar to the results of this study: alpha values for children show an analogous tendency from a value higher than 1 to a value closer to 1. On the other hand, our results show a significant number of alpha values below 1. We do not think that this difference is due to the fact that the languages examined

are part of two different families (Romance and Germanic). We would rather think that this difference could derive from a different methodological choice: differently from Baixeries (2013), we do not use a length normalization of samples. To improve our work, we could repeat the alpha estimation by using a text normalization procedure: this could lead to an improvement in comparability (Baixeries, 2013).

Another additional strategy could be to explore the relationship of the inter-related evolution between *alpha* and the complexity of linguistic output: differently from what has been previously done we claim that it would be better to replace MLU (mean length of utterance) with a more fine-grained index of linguistic complexity as the ISC (Szmrecsanyi, 2004). We are currently developing a Python script to obtain the ISC score for every child utterance.

References

Baixeries J., Elvevag B. and Ferrer-i-Cancho R. (2013). *The Evolution of the Exponent of Zipf's Law in Language Ontogeny*. PLoS ONE 8(3): e53227.

Fan J. and Gijbels I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

Ferrer i Cancho R. and Solé R. V. (2001). *The small world of human language*. Proceedings of Royal Society of London B 2260-2265.

Goodman J., Dale P. and Li P. (2008). *Does frequency count? Parental input and the acquisition of vocabulary*. Journal of Child Language, 35(03), 515–531.

Jackson L., Burkholder R., Flinn G. and Coppess E. (2016). *Working with CHAT transcripts in Python*. Technical report TR-2016-02, Department of Computer Science, University of Chicago.

Lestrade S. (2017). *Unzipping Zipf's law*. PlosOne.

Mac Whinney B. (2000). *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Mandelbrot B. (1961). *On the theory of word frequencies and on related markovian models of discourse*. In: Jacobson R, editor, Structure of Language and its Mathematical Aspects, Providence, R. I.: American Mathematical Society. pp.190-219.

Morgenstern A. and Parisse C. (2012). *The Paris Corpus*. French language studies 22. 7- 12. Cambridge University press - https://hdl.handle.net/11403/colaje/v2.4.

Mucciardi M., Pirrotta G., Briglia A. and Sallaberry A. (2021). *Visualizing cluster of words: a graphical approach to grammar acquisition*. Proceedings of 13th Scientific Meeting of the Classification and Data Analysis Group, Firenze University Press, pp.392-395.

Qi P., Zhang Y., Zhang Y., Bolton J. and Manning C. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Association for Computational Linguistics (ACL) System Demonstrations.

Piantadosi S. (2014). *Zipf's word frequency law in natural language: A critical review and future directions*. Psychon Bull Rev.; 21(5): 1112–1130.

Ridley D.R. and Gonzales E. A. (1994). *Zipf's law extended to small samples of adult speech*. Percept Mot Skills. 79(1 Pt 1):153-4.

Szmrecsanyi B. (2004). *On operationalizing syntactic complexity*. JADT 2004.
Vihman M. M. and McCune L. (1994). *When is a word a word?* Journal of Child Language, 21(3), 517–542.
Yamaguchi N. (2018). *What is a representative language sample for word and sound acquisition?* Canadian Journal of Linguistics / Revue canadienne de linguistique, University of Toronto Press 63 (04), pp.667-685.
Zipf G.K. (1949). *Human behaviour and the principle of least effort*. Cambridge (MA), USA: Addison-Wesley.

Stylometric investigations into translationese: The Baby-Sitters Club across languages

Joanna Byszuk[1], Quinn Dombrowski[2]

[1]Institute of Polish Language of the Polish Academy of Sciences – joanna.byszuk@ijp.pan.pl
[2]Stanford University – qad@stanford.edu

Abstract
This paper examines the multi-lingual corpus of translations of the Baby-Sitters Club, a book series by Ann M. Martin, published from 1986 to 2000, into 6 language versions (distinguishing three French versions, next to Italian, Spanish, and Polish translations) with stylometry, a method of quantitative text analysis based on comparison of texts using differences in word (or other features) frequency distributions, and network analysis. To date, stylometric investigations into translations focused mostly on one language pair, sometimes examining a number of versions of the same text (cf. research by Jan Rybicki or Moshe Koppel). Furthermore, their findings pointed to the ambiguous situation of "the Translator's Invisibility", as in some cases translator's stylistic impact was clearly distinguishable, and in others it was completely lost in favor of the thematic similarities between texts. Here, we address a translation problem comparing across several languages the same set of low-prestige books, with two translators for some texts (one translator adapting a previous translation), and adding the complicating factor of ghostwriters for the original texts. The main finding is that the style of both translators and ghostwriters (even in translation) are visible, although not in all settings.

Keywords: stylometry, middle-grade literature, quantitative text analysis, computational text analysis, translation studies

1. Introduction

The Baby-Sitters Club is a series of middle-grade novels written by Ann M. Martin, published from 1986 to 2000, and translated into numerous languages, becoming international bestsellers and serving as an introduction to American culture to adolescents all around the world. This paper is a part of a larger project, *The Data-Sitters Club* (DSC), which set out to create digital corpora of the series and its translations and examine them with various text analysis methods.

We examine the multi-lingual corpus of translations into 6 language versions (distinguishing three French versions, next to Italian, Spanish, and Polish translations) with stylometry, a method of quantitative text analysis based on comparison of texts using differences in word distributions, and network analysis. The novelty of our approach lies primarily in the application of stylometry to a

translation problem comparing across several languages the same set of low-prestige books, with two translators for some texts (one translator adapting a previous translation), and adding the complicating factor of ghostwriters for the original texts.

The translator's impact on the final work has long been of interest to scholars, particularly after the publication of "The Translator's Invisibility" by Lawrence Venuti. The titular concept he introduced therein has been the cause of much debate, and despite criticism by some translation scholars, it is now considered a core element of translation studies curricula. In his book, Venuti explains that:

> A translated text (...) is judged acceptable by most publishers, reviewers, and readers when it reads fluently, when the absence of any linguistic or stylistic peculiarities makes it seem transparent, giving the appearance that it reflects the foreign writer's personality or intention or the essential meaning of the foreign text—the appearance, in other words, that the translation is not in fact a translation, but the "original." The illusion of transparency is an effect of fluent discourse, of the translator's effort to insure easy readability by adhering to current usage, maintaining continuous syntax, fixing a precise meaning (...) this illusory effect conceals the numerous conditions under which the translation is made, starting with the translator's crucial intervention in the foreign text. The more fluent the translation, the more invisible the translator, and, presumably, the more visible the writer or meaning of the foreign text. (Venuti 1995: 1-2)

The concept of authorship of the translation is a controversial one, on multiple levels. In the prosaic economic sense, until recently, translators have rarely been recognized on book covers or mentioned in book promotions, and have often struggled to obtain copyright (cf. ibidem: 8-9 and further) or significant profit. Furthermore, on an intellectual level, they have not necessarily been considered creators. Venuti references Willard Trask's analogy of translators playacting as authors – "Translators are very much aware that any sense of authorial presence in a translation is an illusion, an effect of transparent discourse, comparable to a 'stunt,' but they nonetheless assert that they participate in a 'psychological' relationship with the author in which they repress their own 'personality'" (ibidem: 7-8). In some sense, the translator's relationship to a text is not dissimilar to that of a ghostwriter: both translator and ghostwriter are integral to the text coming into existence, but their role is upstaged by the notional author, who is treated as the authentic source of the work.

Nonetheless, the visibility of translators' stylistic impact has long been of interest to authorship attribution studies. Jan Rybicki (who has extensively addressed the topic) and his various collaborators have found that contrary to the supposed "invisibility", in numerous cases (e.g. Rybicki and Heydel 2013) the translator can be identified quite clearly, although distinguishing the translator signal is much more difficult than in the case of writers (e.g. Rybicki 2018).

2. Dataset

Our corpus consists of 142 translations of books from the Baby-Sitters Club series credited to Ann M. Martin, along with the original English text for each of these translations. A complete list is included in the appendix[1]. This corpus of translations includes the work of at least 23 translators, and the English originals can be traced to 9 unique authors. While the entire series is notionally credited to Ann M. Martin, a formulaic acknowledgement of someone's help "preparing the manuscript" explicitly indicates the true author of the original text. It is noteworthy that Martin has taken steps to reclaim her copyright from the publisher only for those books where she was the sole author, essentially acknowledging the authorial role of the ghostwriters (Yuvaraj 2022).

The earliest translations in our corpus are from Spain and date to 1988, where both Spanish and Catalan translations were published in parallel, with a number of books translated into both languages by the same translator, Josephina Caball. By 1990, the series was well established in the US with 30+ books, and the range of translations expanded, with separate French translations appearing based out of Belgium and Quebec (the latter to ensure the series' availability across the entirety of the Canadian market). The Polish translations debuted in 1993. Of these early translation efforts, only the Spanish and Quebecois translations were highly successful, with more than 30 books translated into Spanish (over 20 into Catalan), and more than 60 into Quebecois. Only three books were translated into Polish, with an additional two announced but never published. The Belgian translation effort faltered after 16 books, but served as the foundation for a renewed translation effort based in France, starting in 1997. The Italian translation was also part of a later translation wave, beginning in 1999. It was highly successful, with over 70 books translated.

By the mid-2000's, the American series and its spin-offs had come to an end, and new translations had stopped. A graphic novel reimagining of the series by Raina Teglemeier, reinvigorated the series, which began to be re-released in translation in the late 2010's and early 2020's, further fueled by the development of a Netflix series. The approach to these re-publications varied by country: in Poland, an entirely new translation was made, which now continues beyond the three-book scope of the original translation. In Italy, the original books were re-released with only small corrections. In France, the 2000's translations were modified to make the language more colloquial and compatible with changes in the educational standard, however it is unclear who did the work. Thus, while the books marked as "new" French translations are attributed to the same translator as the "old" French translations, they include the impact of an unknown hand.

[1] https://github.com/JoannaBy/BSC-translationese

3. Method – Stylometry

Stylometry involves extraction of textual features and comparing their use across the corpus with a set of statistical methods for classification. It is often applied to authorship attribution as well as distant reading, as it proves useful in discovering patterns and connections between texts, which are otherwise difficult or impossible to distinguish. To this date, stylometric investigations into translations have focused mostly on one language pair, sometimes examining a number of versions of the same text (cf. research by Rybicki or Koppel), and gave no definite answer on "the translator's invisibility", as in some cases translator's stylistic impact was clearly distinguishable, and in others it was completely lost in favor of the thematic similarities between texts.

We resorted to the simplest stylometric procedures, largely following Rybicki's methodology. We use five corpora: 1) English as a baseline for the relations between texts, 2) French, considering all three language variants together, 3) Italian, 4) Polish, and 5) Spanish. For each corpus we conduct a series of cluster analysis classifications, followed by a bootstrap consensus network classification, a method of averaging outcomes obtained with varying range of features to determine which groupings occur with likelihood exceeding the set threshold (in our case: 50%). Following golden rules of stylometry, we use a range of 100-1000 most frequent words as features, and use Cosine Delta (Evert *et al*. 2009, 2015) and Burrows's Delta measure in calculating distances between texts[2]. Finally, we also visualize bootstrap consensus network results in Gephi in order to more easily analyze their results, and further apply a Modularity method of community detection (cf. Newman 2006), a recognized way to find fine-grained connections between nodes that might otherwise skip the observer's eye.

4. Results

All stylometric tests pointed to strong differentiation within the corpus of original works, with easily observable communities in the network (see Fig.1). The reason for this became evident once we added information about ghostwriters – the largest, violet-colored grouping corresponds to works attributed to Ann M. Martin, and other colors to her ghostwriters. Interestingly, early works of Ellen Miles group together with Martin's, which might indicate that in early stages of their collaboration she was highly influenced (or much more likely – highly edited) by the series author. The fact that various ghostwriters have their distinct voices confirms a well-known stylometric truth – mimicking someone's style is difficult and usually impossible to fully obtain.

[2] Full results are available in the appendix on github, here are presented the most important ones.

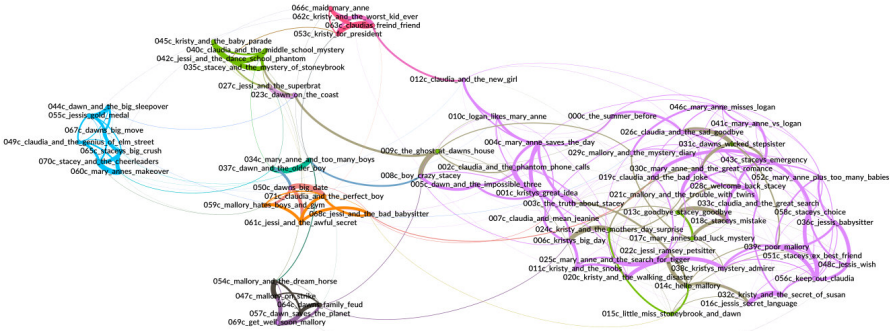


Figure 1. Network visualization of bootstrap consensus analysis of relations between 72 books in their original English text. 100-1000 most frequent words (MFW) as features, colors distinguish between (ghost –)writers.

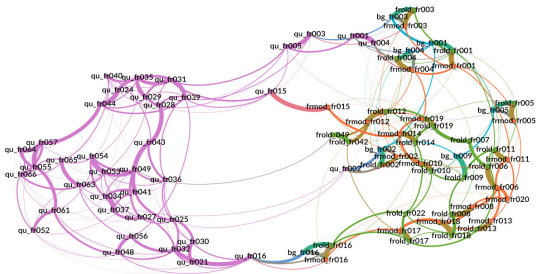


Figure 2. Network visualization of bootstrap consensus analysis of relations between 84 translations to French. 100-1000 most frequent words (MFW) as features, colors distinguish between language variants (also marked in the file name).

The French corpus is the most extensive one we use – it consists of 84 books, and three language variants, as there were separate publications in Belgium, France and Quebec province. Given our fear that Quebecois dialect would be the strongest determinant of the differences, we also run additional tests with the use of 50 and 70% culling (omitting words that do not appear in at least 50% or 70% of the texts in the corpus). However, even after culling the separation was there. Quite expectedly, the most significant difference is between the Quebecois and the European editions. However, more fine-grained detected patterns point to a number of novel observations requiring further analysis. The Belgian translations almost always pair up closely with their French equivalents, which is explainable – the Belgian translation is hyper-localized (names, foods, etc.), but still sets the girls in a suburb of Paris. The first few of the France-French books were "co-translated" – adapted from the Belgian translation. Quebecois translations remain stylistically distant from the remaining corpus. Both European and Quebecois

translations form close clusters indicating thematic or translatorial similarity, and as they are not alike (the clusters in each do not consist of the same texts), this strongly suggests the differences are more rooted in translator's choices and the book content.

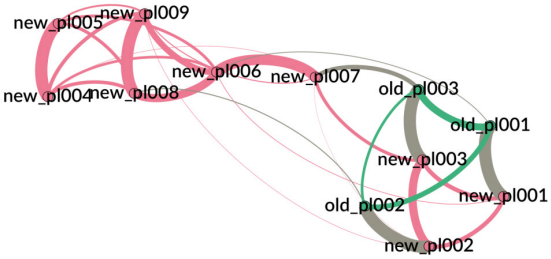


Figure 3. Network visualization of bootstrap consensus analysis of relations between 12 translations to Polish. 100-1000 MFW as features, colors and name distinguish between 90s and 2020s.

The Polish corpus so far consists of 12 texts, including 9 translations done after 2018 (called 'new' here), and 3 translations that were published in the 90s. Stylometrically, Polish translations appear to contribute to the idea of translator's invisibility: the strongest relations are those between two translations of the same novel – despite different translators and close to 30 years distance between them. Furthermore, given that all new translations were done by the same person, it is somewhat surprising that her translations of the first three books are stylistically quite dissimilar to her style in further books. This can be interpreted in various ways, and most likely is due to similar translation strategy – while old translations include some domestications of place names or explanations of concepts and objects foreign to the 90s' reader, Polish translation school is relatively formalized and oriented towards invisibility. It is also noteworthy that stylistical closeness in her further works seems to go in pairs – book 4 close to 5, 6 to 7, 8 to 9. As this links are not detectable in originals, they may result from working on the material in a short time – the new translations are published fast, including the first six books in the span of half a year, and the next ones released every 3-6 months.

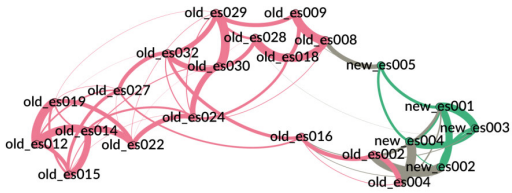


Figure 4. BCA Network visualization of relations between 22 translations to Spanish. 100-1000 MFW as features, colors and names distinguish between 90s and 2020s.

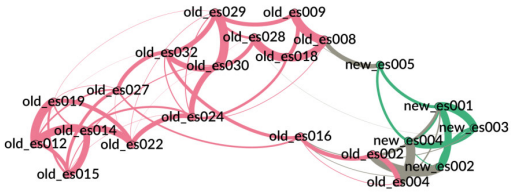


Figure 5. BCA Network visualization of relations between 24 translations to Italian. 100-1000 MFW as features, colors and name distinguish between 90s and 2020s.

The Spanish corpus consists of 22 texts, including five newly published translations. Our corpus includes 3 texts by Josephina Caball (002, 004, 016), with the remaining ones by Conchita Peraire del Molino. Caball's translations are stylistically distinct – evident both to the naked eye and in community detection. Peraire del Molino's translations also form several distinct clusters – 1) 012, 014, 015, 019, 022, then 2) 008, 009, 018, 028, and finally 3) 029, 030 and 032, with 024 and 027 (which were both ghostwritten) joing communities depending on the settings in the analysis. There is no obvious link, with the exception that cluster 2) includes books narrated by Stacey (with exception of 009 by Dawn).

The Italian corpus consists of 24 books, of which 4 are newly published in the 2020's – to the best of our knowledge, they are re-released earlier translations by Adriana Giussani from the 2000s. The corpus also includes her old translation of book 003 (the only one with two translations), and the remaining translations are by her (006, 010, 011), Raffaella Belletti (018, 032), Antonella Baldo (025, 026) and Marina Impallomeni (037 and up), all published in early 2000s. Giussani's translations are consistently clustering together, with the exception of 006 and 011 – which cluster with Belletti and Baldo. When using Classic Burrows's Delta as a distance measure, the texts by these 3 translators form 3 clusters, but with Cosine Delta just two – Giussani's works and her with other authors. Impallomeni's texts also form 2 or 3 distinct groupings: most distinct 1) 047, 057 and 069, and 2) 037, 040, 046, 071, and 3) 049, 055, 062, 066, 067, which in some settings merge into one cluster. This separation is actually quite striking – the clusters correspond to particular ghostwriters, which would indicate Impallomeni successfully mimics some features of their style.

5. Conclusions

As this is a pilot study, we do not intend to make final judgements but to further verify our hypothesis that translators are visible as we extend our corpora. Our study has shown that there is no such thing as one expected stylometric trace of a translator. Across our language collections we saw examples of translators who

are clearly distinguishing themselves from others, and such whose stylistic impact indeed appears "invisible" as connections between their texts resemble those of the original author – even if said author is a ghostwriter working under series' author guidelines, so already someone "playacting as an author". While this observation may seem disappointing, we must emphasize that also traditionally understood authors tend to have a differing strength of their authorial "signal".

For Spanish and Italian we found that while old versus new edition was the strongest differentiator, particular translators's style was very distinguishable. We cannot say the same about Polish, in which case the content determines similarity (which is in line with the results of Rybicki's research on Polish Tolkien's translations and observations on the (in-)visibility of Polish translators). This is also true of French, for which the texts group largely according to the language variant and intense translatorial interventions, such as domestication of proper names. However, French is also the least reliable of our corpora in terms of providing information on particular translators and includes reusing translations from other coutries with changes made by several anonymous editors, therefore it is not unlikely that their signal only seems to be lost because we lack sufficient information.

We plan to further extend the study with more languages, more texts from the series, as well as more points of reference, such as other books from the same genre, similar works by the same group of translators, and to further analyze the texts with more quantitative methods. Many mysteries remain to unfold in this dataset, and we intend to address further methodological questions, such as: is Baby-Sitters Club unique in its visibility of (ghost –)writers and translators or can similar patterns be found among other middle-school series? Is it possible to model different translation strategies and is tranlatorial visibility tied to genre or language tradition?

Acknowledgments

JB was supported by the PL National Science Centre funded project SONATA-BIS 2017/26/E/HS2/01019.

References

Eder M. (2017). Visualization in stylometry: cluster analysis using networks. Digital Scholarship in the Humanities, 32(1): 50-64

Eder M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. R Journal, 16(1): 107-121, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Evert S., Proisl T., Vitt T., Schöch C., Jannidis F., and Pielström S. (2015). Towards a Better Understanding of Burrows's Delta in Literary Authorship Attribution. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 79–

88.
Lodge (2010): https://www.publishersweekly.com/pw/by-topic/childrens/childrens-book-news/article/41524-the-baby-sitters-club-to-reconvene.html
Koppel M. and Ordan N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, USA, 1318–1326
Newman M. EJ. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences* 103 (23): 8577–82.
Rybicki J. and Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish. *Literary and Linguistic Computing*, 28(4): 708-17
Rybicki J. (2012). The great mystery of the (almost) invisible translator: stylometry in translation. In M. Oakley and M. Ji (eds.), Q*uantitative Methods in Corpus-Based Translation Studies*. Amsterdam: John Benjamins, pp. 231-248
Rybicki J. (2009). Translation and Delta revisited: when we read translations, is it the author or the translator that we really read? *Digital Humanities 2009: Conference Abstracts*, pp. 245-47.
Venuti L. (2008). *The Translator's Invisibility: A History of Translation* (2nd ed.). Abingdon, Oxon, U.K.: Routledge
Yuvaraj J. *et al* (2020). U.S. Copyright Termination Notices 1977–2020: Introducing New Datasets. *Journal of Empirical Legal Studies*, February 27, 2022. https://doi.org/10.1111/jels.12310.

Analisi comparativa dei rapporti nazionali: il caso del Progetto Erasmus+ ECOLHE

Stefania Capogna[1], Giulia Cecchini[2], Maria Chiara De Angelis[1], Francesca Greco[2], Emanuela Proietti[2]

[1]Link Campus University– s.capogna@unilink.it mc.deangelis@unilink.it
[2]University of Roma Tre – giulia.cecchini@uniroma3.it francesca.greco@uniroma3.it emanuela.proietti@uniroma3.it

Abstract
In the European projects, there is often a phase in which the results are compared with qualitative, quantitative, or mixed methods. Among the various elements that organize the comparison, an important component is the cultural dimension, since it organizes social actors' practices, which is often carried out with qualitative methods. However, in line with the literature, this dimension is detectable through text mining methods since it determines the choice and association of the words used to organize communication. This work proposes a text mining procedure useful for comparing the documents' symbolic-cultural categories. In particular, Emotional Text Mining was used to study the cultural differences of digital development in higher education among countries through the analysis of the country partners' report of the Erasmus+ Project ECOLHE (Empower Competences for On life Learning in Higher Education), to identify the symbolic-cultural categories and the representations of digital development. Results have important implications for the identification of digital culture development indicators starting from texts, an aspect that could be considered relevant for policy makers in the context of Erasmus+ projects.

Keywords: digital divide, higher education, Erasmus+, Emotional Text Mining, comparative analysis.

Riassunto
Nei progetti europei, spesso, i risultati ottenuti nei paesi partner vengono confrontati con metodi qualitativi, quantitativi o misti. Tra i vari elementi che organizzano il confronto, una componente importante è la dimensione culturale, poiché essa organizza le pratiche degli attori sociali, che spesso viene effettuata con metodi qualitativi. Tuttavia, in linea con la letteratura, questa dimensione è rilevabile attraverso metodi di text mining poiché determina la scelta e l'associazione delle parole utilizzate per organizzare la comunicazione. Questo lavoro propone una procedura di text mining per confrontare le categorie simbolico-culturali presenti nei documenti. In particolare, il metodo dell'Emotional Text Mining è stato utilizzato per studiare le differenze culturali dello

sviluppo digitale nell'istruzione superiore attraverso l'analisi dei rapporti di ricerca dei paesi partner del progetto Erasmus+ ECOLHE (Empower Competences for On life Learning in Higher Education), per identificare le categorie simbolico-culturali e le rappresentazioni dello sviluppo digitale. I risultati hanno importanti implicazioni per l'identificazione di indicatori di sviluppo della cultura digitale a partire dai testi, aspetto che potrebbe essere considerato rilevante per i *policy makers* nell'ambito dei progetti Erasmus+.

Parole chiave: divario digitale, istruzione superiore, Erasmus+, Emotional Text Mining, analisi comparativa.

1. Introduzione

Nei progetti europei è spesso presente una fase in cui i risultati ottenuti vengono comparati con metodi qualitativi, quantitativi o mixed. Tra i vari elementi che organizzano il confronto una componente importante è la dimensione culturale, poiché essa organizza le prassi degli attori sociali, che spesso viene effettuata con metodi qualitativi. Tuttavia, in linea con la letteratura (e.g., Reinert, 1995; Carli and Paniccia, 2002; Greco, 2016), tale dimensione è rilevabile attraverso l'analisi testuale, poiché essa determina la scelta e l'associazione delle parole utilizzate per organizzare la comunicazione. Questo lavoro propone, quindi, una procedura di text mining per confrontare le categorie simbolico-culturali presenti nei documenti che rendono conto dei risultati ottenuti attraverso la realizzazione dei progetti europei.

In particolare, nell'ambito dell'Erasmus+ Project ECOLHE (Empower Competences for On life Learning in Higher Education), nel corso del 2021 è stata condotta una ricerca che, tra i diversi obiettivi, ha avuto quello di indagare il livello di sviluppo delle pratiche di formazione e apprendimento digitale nell'istruzione superiore dei cinque paesi coinvolti nella ricerca. Il tema oggetto di studio attualmente è di grande rilevanza per un cospicuo aumento del ricorso alla formazione *online* connesso con le misure restrittive tese a ridurre il rischio di contagio imposte dalla pandemia. I risultati del progetto, peraltro, hanno evidenziato come il livello di trasformazione digitale nazionale influisca sulla capacità di far fronte anche alle esigenze formative del Paese. La pandemia, infatti, ha portato in primo piano e reso urgente il tema dello sviluppo digitale, caro alla Comunità Europea da più di venti anni.

Attualmente, tutti i partner del progetto ECOLHE hanno redatto un rapporto nazionale in lingua inglese in cui vengono riportati i risultati dell'analisi delle politiche e delle pratiche dello sviluppo digitale dell'istruzione superiore a livello nazionale, istituzionale e professionale, e la loro adozione nel processo di apprendimento. Tutti i documenti hanno la stessa struttura e sono stati sottoposti a un'analisi comparativa di tipo qualitativo per identificare le similitudini e le differenze che caratterizzano i Paesi. I risultati dell'analisi suggeriscono come

uno degli aspetti che influisce sul livello di sviluppo digitale sia la capacità critica, una dimensione che rileva delle categorie che si utilizzano per interpretare la realtà. Secondo una prospettiva socio-costruttivista, queste categorie influiscono anche sulla comunicazione e, in particolare, sulla scelta e l'associazione delle parole. Di conseguenza si può supporre che i report nazionali prodotti per il progetto europeo dai paesi partner siano uguali per struttura, in quanto imposta, ma differiscono nei profili lessicali indicativi degli assetti culturali che hanno organizzato lo sviluppo digitale del Paese.

2. Dati e Metodi

Per comprendere se vi siano delle specificità nelle categorie simbolico-culturali utilizzate per guardare allo sviluppo digitale nell'istruzione superiore, i rapporti prodotti dai cinque paesi partner del progetto europeo sono stati sottoposti a tecniche di text mining. A tal fine i documenti sono stati raccolti in un corpus di medie dimensioni (token = 65.804) con una buona ricchezza lessicale (type/token ratio = 0,09; hapax = 41,8%) e sono state aggiunte delle variabili illustrative per quanto concerne il Paese (Finlandia, Grecia, Irlanda, Italia, Spagna) e il livello di analisi a cui i testi fanno riferimento (nazionale, istituzionale, professionale). Per effettuare l'analisi comparativa si è scelto di utilizzare l'Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020) perché questo metodo si basa su un approccio socio-costruttivista che, sulla base dell'associazione delle parole, individua le tematiche generali e le categorie simbolico-culturali di costruzione di senso. L'ETM si basa su una procedura non supervisionata che, dopo aver effettuato una selezione dei termini sulla base di indicatori lessicometrici (Greco, 2016), effettua un'analisi multivariata che consente di classificare i testi in relazione alla co-occorrenza delle parole.

Inizialmente, i testi sono stati pretrattati eliminando le parti grafiche, i link, le e-mail e i numeri mantenendo solo il testo in minuscolo. Successivamente, sono stati calcolati tre indicatori lessicali per valutare la possibilità di processare statisticamente i dati: i token, il type-token ratio e la percentuale di hapax (Giuliano e La Rocca, 2010). Il corpus è stato suddiviso in 1910 frammenti di testo (T-Lab versione 2018) e sono state rimosse le parole vuote, quelle del rango alto di frequenza e quelle del rango basso di frequenza fino ad un numero di occorrenze pari al numero di report in analisi (n=5) (Cordella *et al*., 2014). Sulla matrice termini-unità di contesto è stata effettuata un'analisi dei cluster con un algoritmo bisecting *k*-means, limitato a 10 partizioni, con l'esclusione dei testi che non presentavano almeno una co-occorrenza, utilizzando il coseno come misura di similarità (Steinbach *et al*., 2000), per individuare le rappresentazioni dello sviluppo digitale nell'istruzione superiore. Il Calinski-Harabasz, il Davies-Bouldin e il coefficiente di correlazione intraclasse sono stati calcolati per individuare la partizione ottimale e sulla matrice termini-cluster è stata effettuata

una analisi delle corrispondenze (Lebart e Salem, 1994) per identificare le categorie simbolico-culturali.
Al fine di facilitare l'interpretazione dei risultati dell'analisi delle corrispondenze, si è assegnato ogni termine ad un solo fattore in ragione del suo contributo assoluto, vale a dire che il termine viene escluso ai fini dell'interpretazione dalla lista delle parole del fattore in cui assume un valore inferiore a quello massimo, in tal modo ogni fattore è caratterizzato da termini specifici e diversi da quelli degli altri fattori.
Infine, sulle tabelle di contingenza cluster-paese e cluster-livello di analisi sono stati calcolati il $\chi 2$ e i residui standardizzati per valutare la differenze nel trattare lo sviluppo digitale tra i paesi e in relazione al livello di analisi.

3. Risultati

I 659 termini selezionati hanno consentito di classificare il 96,5% delle unità di contesto e di individuare sei rappresentazioni (cluster) dello sviluppo digitale che si collocano in uno spazio di senso a 5 fattori riassunto in tabella 1.

Tabella 1. Quadro sintetico dell'interpretazione dei risultati dell'ETM (tra parentesi sotto il fattore è riportata la percentuale di inerzia spiegata e sotto l'etichetta dell'asse fattoriale il valore della coordinata del centroide del cluster)

Cl	*CU %*	*Etichetta*	*Fattore 1*	*Fattore 2*	*Fattore 3*	*Fattore 4*	*Fattore 5*
			34,7%	*19,9%*	*18,6%*	*14,3%*	*12,6%*
			Attori	*Prospettiva*	*Nazionale*	*Digitale*	*Attività IIS*
1	24,1	Impatto dell'Innovazione Digitale	Persone 0,50	0,00	0,00	Struttura 0,58	-0,06
2	14,05	Strategia di Innovazione del Digitale	Istituzioni -0,31	Internazionale 0,25	Vision -0,77	-0,06	Normalità 0,32
3	12,9	Processo di Apprendimento Digitale	Persone 0,52	Nazionale -0,26	Standards 0,45	Processo -0,33	Normalità 0,63
4	17,5	Innovazione Digitale IIS	Istituzioni -0,75	Nazionale -0,76	0,04	0,03	-0,16
5	17,9	Impatto della Pandemia	Persone 0,50	0,07	-0,11	Processo -0,46	Pandemia -0,46
6	13,6	Standard della Qualità	Istituzioni -0,68	Internazionale 0,67	Standards 0,53	0,01	-0,05

I risultati dell'analisi delle corrispondenze evidenziano come i report organizzano la comunicazione entro cinque assi simbolico-culturali: gli attori coinvolti nel processo (persone o istituzioni), la prospettiva (nazionale o internazionale), le dimensioni rilevanti a livello nazionale (la vision o gli standards), il considerare la questione del digitale dal punto di vista delle strutture o dei processi e, infine, se considerare o meno l'impatto della pandemia nel guardare all'attività svolta dagli istituti di istruzione superiore (IIS). Entro queste direttive di senso (spazio fattoriale), si collocano le sei rappresentazioni (cluster) dello sviluppo digitale : l'impatto dell'innovazione, la strategia di innovazione del digitale a livello

nazionale, il processo di apprendimento, l'innovazione digitale degli IIS, l'impatto della pandemia e gli standard della qualità (figura 1).

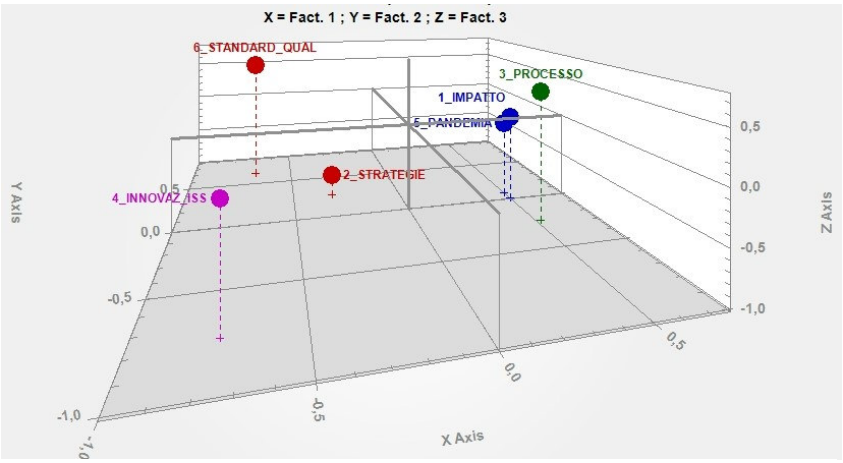


Figura 1. Spazio fattoriale (i primi tre fattori spiegano il 73% dell'inerzia)

I report dei paesi partner del progetto si differenziano significativamente per la presenza o l'assenza di specifici temi (cluster) ($\chi 2$, df= 20, p>0,01) così come si focalizzano maggiormente su specifici livelli di analisi ($\chi 2$, df= 10, p>0,01) (tabella 2).

Tabella 2. Residui standardizzati dell'associazione dei cluster per livello di analisi e Paese in ordine decrescente di DESI
(i valori compresi tra -1,96 e 1,96 non sono stati riportati)

	Livello di analisi			*Paese*				
Cluster	*Nazion*	*Intituz*	*Profess*	*Finlandia*	*Irlanda*	*Spagna*	*Italia*	*Grecia*
Impatto dell'Innovazione Dig.	-8,0	-4,7	10,8		-3,9	7,1		
Strategia di Innovazione Dig.	8,2	-2,5	-5,9	6,3	4,4	-3,7		-4,5
Processo di Apprendim. Dig	-5,5		4,3		-5,6		6,6	-2,5
Innovazione Digitale IIS	6,5	3,6	-8,6	-3,1	-4,7	-3,4		8,5
Impatto della Pandemia	-4,2		3,6	-2,1	13,2		-4,3	-4,1
Standard della Qualità	5,3	2,8	-7,0		-3,6			4,4

I temi dell'impatto dell'innovazione digitale, dell'impatto della pandemia e del processo di apprendimento digitale caratterizzano l'analisi a livello professionale,

mentre quelli degli standard di qualità e dell'innovazione digitale degli IIS contraddistinguono il livello di analisi nazionale e istituzionale.
Il tema della strategia di innovazione digitale invece è specifico solo dell'analisi a livello nazionale, a segnalare la rilevanza di avere una *vision* se si vuole promuovere lo sviluppo digitale di un Paese. Questo tema, infatti, caratterizza significativamente i paesi con il *Digital Economy and Society Index* (DESI) più elevato come la Finlandia e l'Irlanda (EC, 2021), paesi che invece trascurano quello dell'innovazione digitale degli IIS, tema significativo per la Grecia che ha il DESI più basso, probabilmente perché il tema è assente quando il paese ha raggiunto un livello di sviluppo soddisfacente ma assume rilevanza per un Paese che deve ancora promuovere pienamente il proprio sviluppo digitale.
La Spagna, che presenta un DESI superiore alla media europea, si caratterizza per l'attenzione al tema dell'impatto dell'innovazione digitale probabilmente perché, a differenza della Finlandia e dell'Irlanda, il partner coinvolto nello studio è un'università digitale. Infine, i paesi con un DESI al di sotto della media sembrano prestare maggiore attenzione ai processi di apprendimento digitale, nel caso dell'Italia, e all'innovazione digitale degli IIS e agli standards di qualità, nel caso della Grecia. Due temi che, se considerati nel loro insieme, suggeriscono la necessità di creare le condizioni per uno sviluppo digitale del paese che prenda le mosse dall'accademia, punto più alto delle conoscenze scientifiche a livello nazionale, stabilendo dei punti di riferimento come gli standards di qualità, per guidarne il processo e per monitorarlo.

4. Discussione e conclusione

L'utilizzo dell'analisi testuale ha consentito di individuare delle differenze nei profili lessicali nei report nazionali prodotti per il progetto europeo dai paesi partner. L'interesse a individuare tali profili nasce dalla loro connessione con gli assetti culturali che organizzano l'agire degli attori sociali e, di conseguenza, di comprendere in che modo il tema dello sviluppo digitale viene concettualizzato ed espresso a livello nazionale. Possiamo infatti pensare che i ricercatori che hanno redatto i documenti, oltre a essere persone competenti in materia, sono anche l'espressione della cultura della digitalizzazione del proprio paese. Infatti, nonostante i documenti siano stati redatti da esperti di sviluppo digitale, impegnati a promuovere la sua implementazione a livello nazionale, il loro modo di trattare il tema differisce in funzione del livello di sviluppo socio-economico (DESI).
Come evidenziato in letteratura (e.g., Grasso *et al*., 2016), l'assetto simbolico-culturale di una persona può essere ricondotto a tre dimensioni generative dello stesso: una legata all'esperienza personale, che è unica, specifica ed individuale; una connessa al proprio ruolo professionale, espressione dell'appartenenza ad un gruppo sociale con uno specifico percorso formativo e lavorativo e una

relativa al contesto culturale, appresa col processo di socializzazione e ricorsivamente costruita entro le pratiche quotidiane. In questo senso, i risultati hanno importanti implicazioni relativamente alla possibilità di individuare degli indicatori di sviluppo della cultura digitale a partire dai testi, un aspetto che potrebbe configurarsi come rilevante per i *policy makers* nell'ambito dei progetti Erasmus+, in cui il cambiamento culturale è un elemento chiave.

Riferimenti bibliografici

Carli R. e Paniccia R.M. (2002). *L'Analisi Emozionale del Testo: Uno strumento psicologico per leggere testi e discorsi*. Franco Angeli.

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.

European Commission (2021). Digital Economy and Society Index (DESI) 2021: Thematic Chapters. https://ec.europa.eu/newsroom/dae/redirection/document/80563.

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

Grasso M., Cordella B. and Pennella (2016). *L' intervento in psicologia clinica*. Carocci

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

Greco F. e 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. e Salem A. (1994). *Statistique Textuelle*. Dunod.

Reinert M. (1995). I mondi lessicali di un corpus di 304 racconti di incubi attraverso il metodo "Alceste". In R. Cipriani and S. Bolasco, editors, *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*. Franco Angeli.

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.

Three main dimensions in the construction of the expert during the Italian Covid-19 vaccination campaign: positional, reputational, and communicational spheres in comparison

Rosanna Cataldo[1], Gabriella Punziano[1], Carmine Barricelli[1], Gabriele Luciano[1], Barbara Saracino[2], Ferdinando Iazzetta[3]

[1]Università degli Studi di Napoli "Federico II" - rosanna.cataldo2@unina.it gabriella.punziano@unina.it c.barricelli@studenti.unina.it gab.luciano@studenti.unina.it
[2]Università di Bologna "Alma Mater Studiorum" - barbara.saracino@unibo.it
[3]Università di Roma "La Sapienza" - ferdinando.iazzetta@uniroma1.it

Abstract
The purpose of this article is to understand what mechanism leads specific actors in the scientific community to assume the central role of experts during the COVID-19 pandemic vaccination campaign. We will try to understand whether the figure of the expert is influenced by his or her institutional role, network of collaborations, and academic network as a proxy for his or her reputation and, which figure emerges instead from mainstream media.

Keywords: visible expert, media representation, COVID-19 vaccination campaign.

1. Introduction

The involvement of scientific experts in media coverage and public exposure during the Covid-19 pandemic recalls Goodell's (1977) concept of "visible experts". Their presence in public communication leads to new changes in the dynamics between science and society (Maasen and Weingart, 2005; Bucchi and Trench, 2014). In time of the Covid-19 pandemic the role of expertise and of the scientist becomes more and more crucial in the academic debate (Algan *et al*., 2021), therefore with the symbolic launch of *VaccineDay* in Italy, we have formulated a research question that has become the guiding principle of this research. It is a question of understanding the mechanism that leads specific actors in the scientific community to take on the role of expert and become central in the political and health agenda to effectively promote, through the mass media, interventions both in support of vaccine decisions and in response to cases of misinformation ready to raise social alarm or to stoking seeds of vaccine hesitancy.

2. Research Methodology and Objective

Our research draws inspiration from the report *A year of Pandemic, 1st report of the TIPS project - Technoscientific Issue in the Public Sphere* that investigates the pandemic through media coverage in newspapers to monitor the quantitative presence and qualitative evolution of technoscience in the public discourse, in this case with the aim of exploring the centrality on the public sphere of experts. The main interesting result in relation to the goals of our work was the rank of sharing of the thirty-three experts[1] in newspapers that somehow implies a media construction of the figure of the expert in this period. Building on this explanation, the central research question that guided our project was aimed at understanding the world of narratives and self-constructed or inferred representations regarding the figure of the expert resulting from the interplay between positional, reputational, and communicative spheres. However, because the Covid period was filled with crucial and decisive moments in the national health agenda and pandemic governance, we decided to focus our work by concentrating the emergence of the figure of the expert in connection with one of the most specific themes, namely that of the vaccination campaign. Fragmenting the main research question, this paper focused on three sub-questions: Is the figure of the expert influenced by the institutional role they play such as their position? Is their media relevance influenced by the academic collaboration and network as proxy of their reputation? What is the social representation and which figure emerging instead from media mainstream like press and social like Facebook and Twitter? To answer these questions, it is necessary to divide the plan of inquiry into three areas that cover the cognitive scope of this paper focusing on the figure of the thirty-three experts. The first dimension is cantered on the positional area that is analysed here from the side of personal and biographical information. The second dimension concerns the area of reputation here operationalized as data regarding public positions, academic esteem, and online influence. The last dimension focuses on the area of communication processed as representation in social and traditional media. For the construction of the empirical documentation[2], the selected experts were

[1] *Walter Ricciardi, Silvio Brusaferro, Anthony Fauci, Andrea Crisanti, Massimo Galli, Roberto Burioni, Giovanni Rezza, Fabrizio Pregliasco, Franco Locatelli, Ilaria Capua, Alberto Zangrillo, Matteo Bassetti, Pierluigi Lopalco, Giuseppe Ippolito, Francesco Vaia, Mike Ryan, Massimo Andreoni, Nicola Magrini, Roberto Cauda, Giorgio Palù, Guido Silvestri, Antonella Viola, Maria, Rosaria Capobianchi, Alberto Mantovani, Giovanni Di Perri, Silvio Garattini ,Giuseppe Remuzzi, Massimo Clementi, Paolo Ascierto, Maria, Van Kerkhove, Eugenio Gaudio, Luca Richeldi, Alberto Villani.*

[2] The tools used for collection were FanpageKarma for social media, Publish or Perish for publications, Google Search for Curriculum Vitae, Volocom for newspapers and social news extracted thanks to a defined set of keywords, while Ucinet, Excel, Gephi, SPSS and T_LAB were used for the analysis.

placed in two matrices: one useful to cover the first two areas that we will investigate with the following data: expert, age, *Tips* sharing on articles, disciplinary field, specialization, possible top institutional role and political appointment in a technical-scientific committee (CTS) – all data functional to apply network analysis paths; the other matrix built to cover the area of communication that was constructed by extracting content from the national newspapers *Il Corriere dell Sera*, *La Repubblica*, *La Stampa* and, from social networks *Facebook* e *Twitter*[3] on which a quantitative content analysis was conducted. Specifically, 7,728 instances including articles and posts/tweets from 01/11/2020 to 30/10/2021.

2.1 Positional area

Socio-demographic characteristics highlight that the list is composed of 4 women, of whom only Maria Van Kerkhove is under 50 years old, and 29 men. So, how is the figure of the expert constructed and accredited? To answer this question starting from the positional area, our proposal was to reconstruct the pathways understood as relevant institutional steps in education and in the professional mode, studying centralities and roles in these specific networks, but also changes and characteristics that make the profile of the expert identifiable. Thanks to the generation of an affiliation matrix, we reconstructed the university educational network, considering both undergraduate and doctoral pathways and the main affiliated institutions (graph 1 and 2). From the results it is possible to trace three large clusters. The first with a predominance of training abroad, the second with the University of Rome "La Sapienza" as the central node closely linked to the Federico II in Naples and, the last, with the centrality of the University of Milan. The analysis allows us to view both the most influential universities for our research, where we can deduce how the geographical proximity between them tends to strengthen collaborative relationships, and, how most of the experts with high equity value have a university connection/affiliation abroad. In the positional area, the figure of the expert is more and more oriented to the plurality of relationships and multidisciplinary in the different networks of institutions that sees them participating and central, on the other hand, figures of experts such as Bassetti, Capua, Ascierto and Fauci always central but more isolated in their own network. In addition, for the positional area the construction of the figure of a scientist becomes, especially in times of health crisis and vaccination campaign, a credential that can be spent on several tables. On the

[3] Keywords: "vaccin* and espert* e covid*, vaccin* and espert* e coronavirus, covid and scientiat* and vaccin*, coronavirus and scientiat* and vaccin*, vaccin* and espert*, vaccin* and scientiat*".

one hand, that of political governance (Beltrame, 2007; Rubin, 2020) where the expert can take on an advisory role, such as the presence within the CTS of the Conte government of 4 experts: Brusaferro, Locatelli, Ippolito and Silvestri – with the confirmation of the first 3 plus the appointment of Rezza and Palù in the CTS of the Draghi government. Among the associations and organizations in which scientists have worked, the most important are the National AIDS Commission (CNA), the Superior Health Council (CSS), the Superior Health Institute (ISS) and the World Health Organization (WHO).

Figure 1. Training network and PHD study Elaboration Ucinet, source online CVs. (Min linkage:1)

Figure 2. Experts Institutions Network Elaboration Ucinet, source online CVs. (Min linkage1)

2.2 Reputational area

The reputational area explores the coverage at the academic level both with the different scientific production and the collaboration network among the various experts. The academic relevance of the experts was reconstructed to consider and compare it with the pre-covid relevance. It was conducted through the automatic calculation of h-index (or Hirsch index, 2005), quantified the prolificacy and scientific impact of the experts of reference thanks to the number of publications and citations. Among the authors with the highest total h-index we find Fauci, Mantovani, Remuzzi, Silvestri and Locatelli, while on topics centred on Covid-19 of the various experts, also present in the CTS, we note the reconfirmation of Fauci among the most productive, unlike for example Silvestri who is committed to follow and work more on pandemic governance. Following the Covid-19 h-index, experts such as Galli, Zangrillo, Bassetti and Pregliasco emerge to exploit the topic scientifically but not only. In fact, this shows how the figure of the visible expert in the era of pandemic has been built not only thanks to the previous path of publications but also by the possibility of doing research using the central theme (Vaia). Furthermore, it is argued in the literature (Collins and Evans 2002) that in science, co-authored publications of scientific articles seem to be among the most important ways to recognize that a person

has made significant contributions to the development of a research field. In our case, the list of reference experts allows us to place among the "collaborative contributors" Zangrillo, Bassetti, Galli and Ippolito as nodes in the network with more publications, while Clementi and Andreoni with a wider network of contacts. In fact, from graph n.3 it is evident that the latter are a real link between experts both in a multidisciplinary sense and as a centrality assumed in the network.

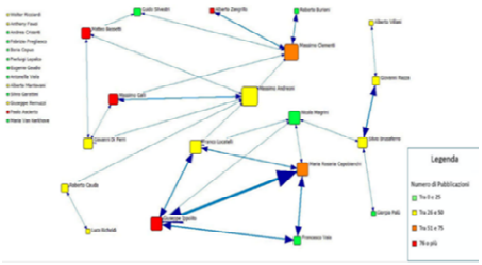


Figure 3. Experts Collaborations' Network -only Covid-19's publications. Elaboration with Ucinet, source online CVs. Minimum linkage: 1 Size by: Degree Centrality. Color by: num of publications

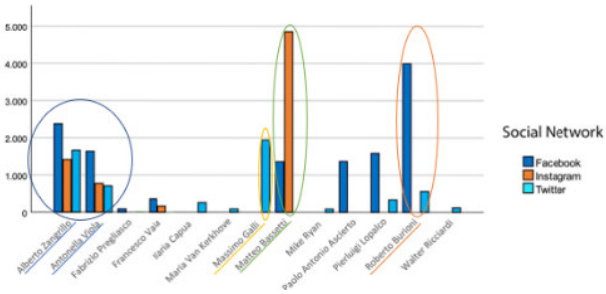


Figure 4. Social Network Usage by Experts

A second element for the positional area is the centrality of the social platform in the construction of the figure of the experts. The shifted scientific debate on the web leads to the question of whether being online is a must for science managers or, again, whether social media feeds institutional distrust and fake news (Bradshaw and Howard, 2018). As we know well from the literature (Colombo, 2015) each platform has its own target audience, and being followed on one of these platforms clarifies the locus of importance within which the individual expert fits. We observe a different distribution of followers among the three reference platforms. Burioni, in fact, on Twitter and Facebook is the first

for followers with more than 100,000 followers, Bassetti (not present on Twitter) the most followed on Instagram and, with the virologist Ilaria Capua alone, keeps pace by exceeding the roof of 100,000 followers. This highlights how, in the positional sphere, presence in mainstream media is reserved for only a few of the experts and only those capable of creating their own language, especially in a format characterized by a feature that sets them apart. At this juncture, it becomes legitimate to project these considerations also by virtue of their presence in television broadcasts. In fact, the most "followed" experts on social networks are precisely those characters that we have seen more times on TV. If we look at the graph, n.4, there are those who use all platforms with the same intensity and tend to have a multi-platform profile of the content they produce, such as Zangrillo and Viola.

There are experts who use more visual and emotionally intense content, like Bassetti who dominates on Instagram, and others who focus on the rational component and public and political debate, like Galli, who makes Twitter his favourite platform. Still, others choose Facebook as their digital forum, partly because it reaches a more generalist and varied audience, and among these Burioni stands out, followed by Lopalco and Ascierto. Yet graph 4 showed us that Burioni is the most followed on Twitter along with Capua, who however does not post much on this space; Bassetti is the most followed on Instagram as well as Burioni on Facebook. This means that the relational component and the public and political debate that passes through Twitter, rather than being dominated by the amount of activity of the participants on that platform, is instead connected to other dimensions of reputation and importance, probably dictated by the content conveyed and not by the way it is conveyed.

2.3 Communicational area

To analyse the characteristics of the communicational area, press was used as a proxy for representation in traditional media and social networks as a proxy for representation via social. Texts were organized into a case matrix for variables such as: media type, date, the complete corpus and, the presence or absence of the reference experts in the corpus with 32 dichotomous variables. From table 1 we can see that social networks cover more than 61% of cases compared to 38% of newspapers.

Table 1. Type of media and Phase

Type of media	First Phase	Second Phase	Third Phase	Total
Newspaper	12,88%	16,63%	8,70%	38,21%
Social network	21,30%	17,85%	22,64%	61,79%
Total	34,18%	34,48%	31,34%	100,00%

In the construction and representation of expert across the two media types we find both similarities and contrasts with respect to the most used keywords. Indeed, the main queries for both media types are vaccine and expert but, nevertheless, there is a tendency to create a representation space that moves away from the specialist domain. The construction of the figure of the expert using keywords is central in daily articles in the first phase and with a further increase in the second and then reduced in the third. While for social networks, if Twitter follows the same line, Facebook with 18.52% concentrates expertise in the last phase. From this we can deduce how the centrality of the expert and the construction of it for the area of communication was dominated by the plurality of content produced on web 2.0, often recalling interviews and products of other media. Subsequently, thanks to quantitative content analysis techniques, it was possible to study the different associations between words by identifying those that appear closest to each other. Using T-Lab it was possible to study and explore the semantic relationships between words within the whole corpus and the network of experts in relation. From the outputs it is possible to notice a different centrality of the figures of the expert and the scientist, differentiating the representation by type of media and elaborating thanks to keywords sequences, elements whose lexical units are present in the corpus or in a subset of it. From graph n.5 representing the construction of the *expert* on the newspapers, we notice associations ranging from considering the period and, therefore, between the temporal distance of the administration and the duration of the vaccine (words such as week, month, years), to explaining in terms of health the efficacy and usefulness (also with the arrival of new variants of Covid-19) and resuming once again the importance of exposing the numbers of the pandemic. Moreover, from graph n.6, for newspapers, it is clear how the construction of the *scientist* inserted in the network returns an interconnection directed both to science (science, pandemic, vaccinate) and to people. Just the high frequency of the plural possessive forms *our* is not a coincidence, in fact, together with world and people puts at the centre of the debate during the vaccination campaign a figure of the scientist inclusive and at the public service.

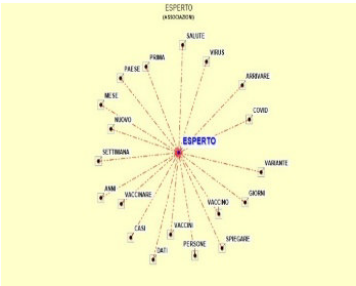


Figure 5. Topic of the expert on newspapers.

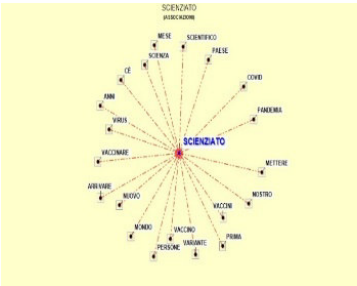


Figure 6. Topic of the scientist on newspaper.

On social networks, the figure of the expert stands out with a different narrative and construction. We find words like Anti, doubt, Astrazeneca that we can explain by assuming that the topic and generic posts on social networks are more varied with a wide spectrum of topics and a lexicon often oriented to a specific type of target. Changing the communicative register also automatically emerges a question/answer to uncertainties and the new (with the word *third* placed at the centre). The lexical patterns in which words like *use*, *anticovid* participate, together with the previous ones show how the participation of the expert in the construction of knowledge for the audience has changed. Even for the figure of the *scientist* on social networks, the words identified are hardly surprising as they are the ones most echoed and known during the pandemic period. However, interesting for interpretation purposes are the distributions of the keyword's *death* next to *effective* ready to recall the usefulness of the vaccine and, the figure of the expert Bucci the only one not present in the reference list and that the research on the corpus reported among the most frequent.

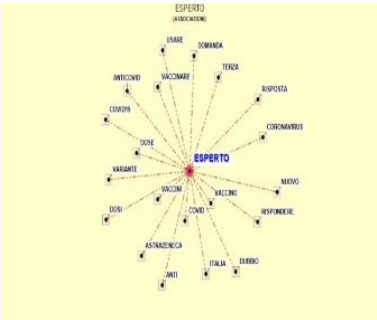


Figure 7. Topic of the expert on social network.

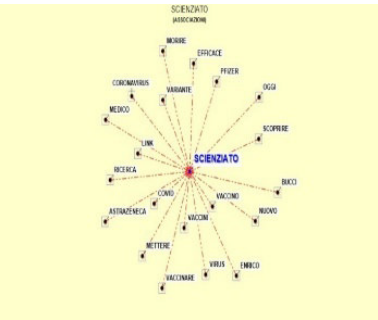


Figure 8. Topic of the scientist on social network.

3. Concluding remarks

In this contribution, the three areas positional, reputational, and communicative, show how the construction of the figure of the expert is dependent on several factors, often interconnected and in interplay. In fact, it is possible to say that the relevance and emergence of the expert in relation to the topic of the vaccine campaign is accelerated both by his scientific activity (consisting of both a dense network of interdisciplinary relationships and with different research institutions) and by the intersection of mainstream and social media. In fact, even within the more traditional media, such as newspapers, there has been a selection of personalities and a construction of the figure of the scientist who has a propensity for a type of communication and dissemination of information characteristic of

digital platforms. However, from the elaboration emerge specific frames and frames of representation of the figure of the *visible expert* as opposed to those *not visible* or at least, *not visible to the media*. Such a view reflects both how the entry of scientists into media programming is not always linked to their academic activity, and how the priorities of experts in an institutionally relevant position are to not communicate. Therefore, what makes a scientist the expert of the moment? This increasingly controversial question can be answered through future analyses on the collected content.

References

Algan Y., Cohen D., Davoine E., Foucault M., and Stantcheva S. (2021). Trust in scientists in times of pandemic: Panel evidence from 12 countries. Proceedings of the National Academy of Sciences, 118(40).

Beltrame L. (2007). Ipse dixit. I premi Nobel come argomento di autorità nella comunicazione pubblica della scienza, «Studi di Sociologia», Vo. 1, pp. 77-98.

Bradshaw S. and Howard P.N. (2018). Sourcing and Automation of Political News and Information over Social Media in the United States, 2016-2018. In Political Communication, 37:2, 173-193.

Bucchi M. and Trench B. (2014). Manuale di Comunicazione Pubblica della Scienza e della Tecnologia (2a edizione). Londra: Routledge.

Colombo F. (2015). Social TV: Produzione, esperienza e valore nell'era digitale. EGEAspa

Collins H. and Evans R. (2002). The Third Wave of Science Studies. In Social Studies of Science 32(2).

Goodell R. (1977). The visible scientists. London: Little, Brown.

Maasen S. and Weingart P. (2005). Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making. Switzerland: Springer.

Rubin A. (2020) Fiducia, reputazione e visibilità di esperti esperti. il caso-studio della notte europea dei ricercatori. In 82- LXIV | 2020. (pp. 25-42).

The Italian Social Mood on Economy Index: recent methodological developments

Elena Catanese, Monica Scannapieco,
Mauro Bruno, Luca Valentino

Istat – catanese@istat.it scannapi@istat.it mbruno@istat.it
valentino@istat.it

Abstract
Istat, since October 2018, publishes every quarter an experimental daily index, i.e. the Social Mood on Economy Index (SMEI). This index is derived from real-time samples of public Italian tweets, containing at least one word belonging to a specific set of filter keywords. The sentiment scores are obtained by means of an Italian sentiment lexicon, namely a vocabulary whose lemmas are associated to pre-computed sentiment scores.
Due to its experimental status, and due to the rapid growth of sentiment analysis methods, two core methodological aspects have been recently analyzed and re-designed: the filter and the vocabulary. The scope was to answer two main questions: to what extent was the set of key words, properly grasping economic conversations; was it available, a different and more recent lexicon, which could improve the quality of the scoring of tweets and therefore also the dynamics of the daily series? To answer the first question a Word Embedding analysis has been carried out. This allowed elaborating a sub-filter. Peaks observed in the sub-filtered series seem to be more fitting to the purpose of the index, i.e. providing a sentiment on the economy. Concerning the vocabulary, a more recent Italian sentiment lexicon, Distributional Polarity Lexicon (DPL), has been taken into account. Comparisons between two series, obtained by using the same number of daily tweets, but different lexicons, show many encouraging results in favor of DPL, namely: (i) more coherent downwards during Covid waves are observed; (ii) the correlation with some economic series increases, and (iii) it is more intuitive to extrapolate the trend.
Overall, the current methodology for the SME index has been revised to embody these findings and improvements.

Keywords: twitter; sentiment analysis; high-frequency series; text mining

1. Introduction

In recent years, the interest for the possible use of Social Media data sources to analyze the public mood on specific topics has been growing and National Statistical Institutes (NSIs) have been investing on methods that could support such purpose, including Sentiment Analysis. The goal of sentiment analysis is to extract the mood of a natural language text; there are methods based on

unsupervised learning techniques, others on lexicon-based methods, and others on combinations of two previous approaches. The Italian National Institute of Statistics (Istat) is at the forefront in this area, indeed it is among the first NSIs to have published an experimental index that measures the daily mood of Italian Twitter users about economy: the Social Mood on Economy Index (SMEI) (Zardetto 2018). SMEI is published[1] since October 2018 in the experimental statistics area and is updated every quarter; the time-series begins in February 2016. Istat adopted an unsupervised approach because supervised methods require large sets of manually labelled texts; more precisely, it adopted a lexicon-base method (more details can be found in the methodological note[2]). Briefly Istat's procedure, consists of three fundamental steps: (i) collection of a daily sample of public tweets matching a filter consisting of relevant keywords inherent to economy; (ii) calculation of tweets' positive and negative sentiment scores, as weighted averages of the sentiment scores of each word in the tweet; (iii) computation of the daily index value as an appropriate central tendency measure of the score distribution of all the tweets. Subject matter experts, borrowing keywords from the Italian Consumer Confidence Survey (CCS) questionnaire[3], designed the filter, used in the first step. Such filter, in the following is referred to as CCS-filter, collects on average about 57,000 tweets per day.

In the present work, we try to understand to what extent this set of key words grasps economic conversations was and - to measure quantitatively the impact of noise in the social mood, by running simulation with different filters. Further, we will show improvements in the quality and interpretability of SMEI time-series by comparing simulations with two different Italian lexicons, namely Sentix and DPL. The major focus is in the period related to Covid-19 pandemic, as the serious economic crisis, witnessed by a collapse of all main macroeconomic variables, was not clearly recorded by the SMEI, but we are also interested in increasing correlation with other economic series in non-pandemic times.

In Section 2, we describe the redesign of the filter and highlight its impact on the time- series. In Section 3, we compare the vocabularies and discuss the improvements brought by DPL. Finally, in Section 4 we summarize the results of the new procedure and the main achievements.

[1] Available at https://www.istat.it/en/archivio/219600

[2] Available at https://www.istat.it/it/files//2018/07/Methodological_Note_social-mood.pdf

[3] It consists of questions deemed most appropriate to assess the optimism/pessimism of consumers (namely: assessments and expectations on the Italian general economic situation, on unemployment, assessments on households' financial situation, opportunity and possibility of savings, current of durable goods purchases, assessments on family budget)

2. Redesign of the filter

Ideally, tweets' filters should be able to capture relevant messages and eliminate off-topic ones. However, some off-topic tweets may pass the filter. Off-topic "viral" tweets, being re-twitted and quoted thousands of times a day, may unduly affect the daily index and introduce bias. SMEI calculation algorithm adopts a special surveillance system that allows making the index robust against these possible contaminations. Although the major risk connected to the inclusion of off-topic tweets is untruly outliers, systematic errors, induced by the wrong definition of the filter, may alter the trend of the time-series or induce some seasonal patterns or calendar effects. Another complication arises from the fact the same word may have different meanings and be used in different contexts. For this reason, we tried to understand, using word embeddings, whether the choice of an expert-based filter was able to grasp relevant information for economic statistical purposes. Word embedding is a technique that permits to quantify and categorize semantic similarities between linguistic items based on item co-occurrences in a very large sample of texts. This analysis allows for recognizing words that on «average» are not used in economic contexts. With respect to the specific technique, we used a word2vec model (Mikolov, 2013), trained on large sample of tweets filtered for SMEI between 2016 and 2018. In Figure 1, we show some examples of iterative graph representations of the word embedding model used to analyze the filter keywords. In Italian several words have multiple meaning such as "*borsa*", bag or stock-market, and "*benessere*", well-being or wealth. It is not clear, a priori, which meaning is more frequent on twitter conversations. The analysis of word embedding graphs pointed out that both terms were used in a meaningless sense for economic analysis purposes. However the use of composite expressions such as "*benessere economico*" (wealth or economic prosperity) or "*perdite in borsa*" (losses in stock-market) ensure the relevance of the filter, as shown in the right graph of Figure 1 (*Unicredit* is a bank, *piazzaffari* is a Mailand's stock market).

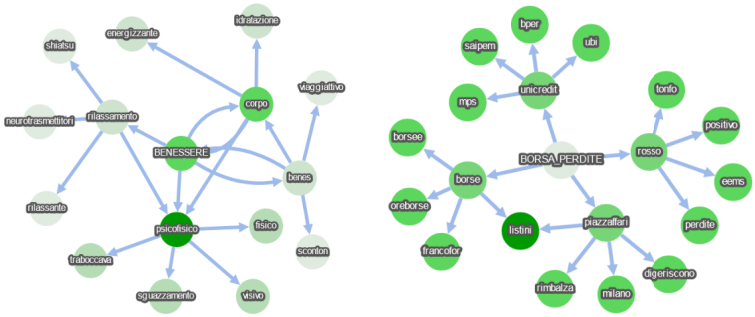


Figure 1. Word embedding graphs for terms : « BENESSERE », « PERDITE in BORSA »

For these reasons, a second-level filter has been added, to exclude out of scope first-level sampled tweets, in the following referred to as Istat-filter. More in detail, Istat-filter contains 115 keywords (words or lemmas) refining CCS-filter by declining some words into more specific composite expressions or by excluding some words. This refinement allowed selecting tweet's samples with more specifically economic content. However, the volume of samples decreases significantly (on average from 57'000 to 26'000).

2.1 Results

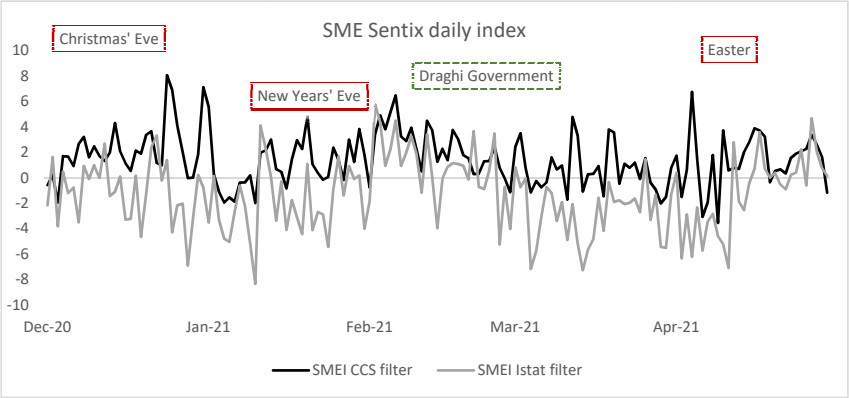


Figure 2. impact of the filter in the SMEI daily series for the period December 2020 - April 2021

The entire time-series has been recalculated using Istat-filter in two different scenarios (different lexicons). Same patterns emerge when comparing the differences induced by the change of filter. In particular, in both settings the Istat-filter daily series and the CCS- filter show a fair correlation, between 0.6 and 0.7 in the different settings (see Figure 2).

It is worth to notice that Christmas and Easter effects visible in CCS series (marked in red in Figure 2), disappear in Istat-filter series. In the new series, such positive peaks disappear, while others, more clearly related to economy, persist (e.g. the nomination of Draghi as prime minister in February 2021). The rationale behind why these series differ, especially for religious holidays has to be linked with the fact that keywords related to the personal climate have been disregarded in the second level-filter.

3. Redesign of the vocabulary

Lexicon based methods, namely vocabularies whose lemmas are associated to pre-computed sentiment scores, have the advantage of being cost-effective.

However they are static, and the lemmas included may depend on the specific training set used. In Italian, at the time being and to the best of our knowledge, there are only two public lexicon available: Sentix (Basile and Nissim, 2013) and DPL (Castellucci *et al*, 2016) both automatically trained on Twitter data. There are drawbacks when using outdated lexicons, for instance the feeling and sentiment associated with a word may change over-time and the lexicon coverage may decrease because of neologisms and new acronyms (e.g. "Covid" or the economic recovery plan "PNRR"). Further, it is not straightforward to compare two lexicons, due to the different number of lemmas and to the different scoring (different sentiment topologies). Sentix integrates several existing, independent lexical resources (WordNet, MultiWordNet, BabelNet, SentiWordNet), therefore it contains many duplicated lemmas (also composite expressions) ~75'000 lemmas overall, only ~42'000 unique. DPL contains two different vocabularies: 75000 natural language words (68000 unique), 143000 unique expressions (in economia, @economia, and so on). We decided to merge these two vocabularies, and the intersection turned out to be equal to 57000 roughly. Since we were not able to identify the context and attribute the different sentiment scoring of same words, to ensure unambiguous and reproducible results, duplicated entries, after text cleaning, were de-duplicated by averaging sentiment scores. The so described obtained vocabulary, consists of almost 139000 entries, and has been utilized in the following simulations. Finally, it must be noted that, the topologies of Sentix and DPL are different. In Sentix only positive and negative scores are available, while in DPL also neutral score is measured. However, it must be noted that the contribution of the neutral score is null in the sentiment scoring procedure.
Since both vocabularies turned out to be outdated with respect to the pandemic, we decided to run our simulations, by adding the word Covid, with the bronchopneumonia score and used the Istat filter previously described.

3.1. Results

We compared the SMEI calculated using the Sentix lexicon and the SMEI calculated using the merge of DPL lexicons with two high-frequency indices related to the health-crisis, namely: (i) Google Trends, a search trends feature that shows how frequently a given search term is entered into Google's search engine relative to the site's total search volume over a given period of time, in particular in Fig. 3a we show the weekly searches in Italy of Covid and Coronavirus; (ii) the Stringency Index (Ritchie *et al*, 2020), a daily measure of limitations due to government interventions to limit the spread of the virus. For the comparisons, we used, respectively, a weekly average (Fig 3a) and a weekly moving average of SME simulations (Fig 3b).

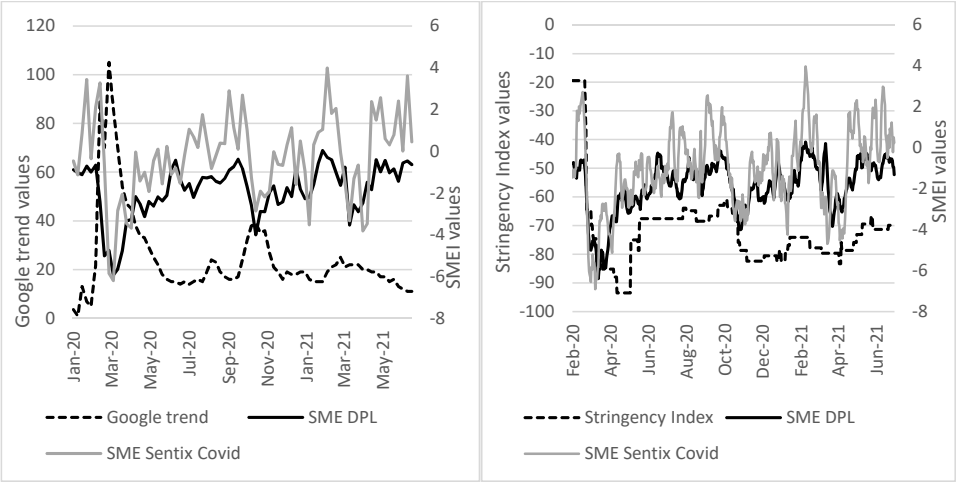


Figure 3. Impact of new lexicons in the SME Series (Sentix Covid, DPL) compared with non traditional high frequency series in the pandemic period: in the left (Figure 3a) the comparison of the week average with Google Trend; in the right (Figure 3b) the comparison with the daily Stringency Index

Even if in the simulations in Figure 3 the two daily SME indexes are obtained from the same sample of tweets, the DPL version shows better correlations with indices related to the pandemic: DPL correlation with Google trend is -0.77 (inverse relation, because high volumes of conversations about Covid are associated with a negative mood) and with Stringency index is 0.55; while Sentix correlation is respectively -0.45 and 0.33.The figure clearly shows that the time-series obtained by DPL lexicon is less volatile and a has more intuitive trend.

4. Conclusions

In the previous sections, we highlighted the enhancements of the SMEI achieved by the redesign of the filter and the change of the vocabulary. Although the procedures and methods have not been altered, the introduced changes have a strong impact on the dynamic of the raw index. On one side, the change of sampled tweets, induced by the changes in the filter, enhanced the economic interpretability of observed peaks. On the other side, some phenomena, such as Draghi Government (February 2021), remain observable confirming the robustness of the procedure. The lexicon change structurally modifies the raw index, by significantly reducing the volatility of the time-series. The introduction of Covid related terms is responsible for the significant breakdown not previously present when using the Sentix vocabulary. Another important tool to validate the

quality of the updated SMEI is to compare the new time series with traditional economic indexes in non-pandemic times (see Figure 4). We chose the Consumer Confidence Index[4] (CCI), a monthly time-series elaborated by the CCS survey. It is not straightforward to compare these two indexes, because the Consumer Confidence Survey collects information within the first 2 weeks of each month and the April CC value is missing. For these reasons, we observe a right shift in the CCI during the first lock down. Although the above-mentioned limits, the correlation of these two monthly series within three years is 0.64 and within the period Jan-19 to Mar-21, it rises to 0.9.

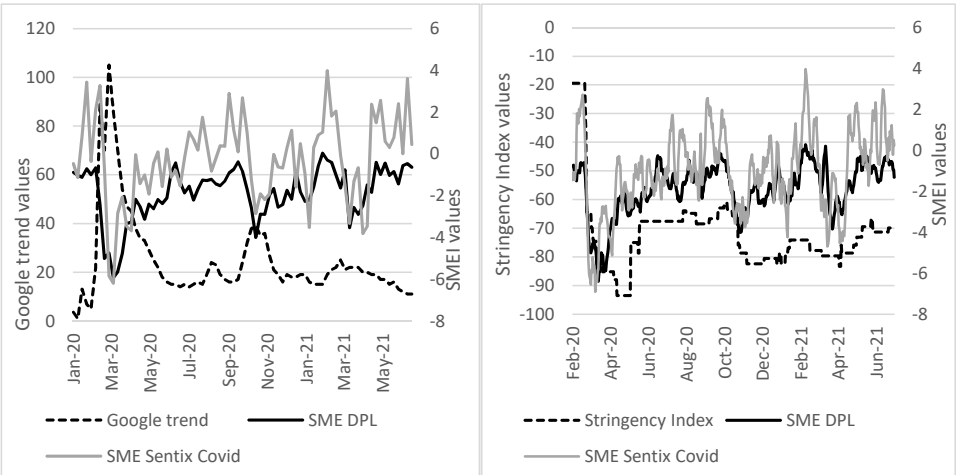


Figure 4. Comparison between the Italian Consumer Confidence Index and the monthly average of the updated SMEI (using DPL lexicon and Istat Filter)

For all these reasons, the entire time series of SMEI has been reviewed and published since the last quarter of 2021 with the new methodology.

References

Basile V. and Nissim, M. (2013). Sentiment analysis on Italian tweets. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 100-107), Atlanta, USA.

Castellucci G., Croce D., Basili R. (2016): A Language Independent Method for

[4] For details see: https://www.istat.it/it/files//2022/03/Business-and-consumer-confidence-MARCH-2022.pdf

Generating Large Scale Polarity Lexicons. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portoroz, Slovenia, 2016.
Mikolov T.; *et al*. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781
Ritchie H., Mathieu E., Rodés-Guirao L., Appel C., Giattino C., Ortiz-Ospina E., Hasell J., Macdonald B., Beltekian D. and Roser M.(2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org.*
Zardetto, D. (2018). Using Twitter Data for the Social Mood on Economy Index, Atti della XIII Conferenza nazionale di statistica, Rome, 4-6 July 2018, ISBN 978-88-458-2016-8, pp. 385-390.

Does an author leave a syntactic footprint?

Radek Čech[1], Miroslav Kubát[1], Ján Mačutek[2,3],
Michaela Koščová[2]

[1]University of Ostrava – cechradek@gmail.com
miroslav.kubat@gmail.com
[2]Mathematical Institute, Slovak Academy of Sciences –
jmacutek@yahoo.com kosc.michaela@gmail.com
[3]Constantine the Philosopher University in Nitra

Abstract
Results of an authorship analysis of Czech texts based on syntactic relations are presented. For the analysis, 82 annual speeches of Czech and Czechoslovak presidents delivered on the occasion of the end of the year were used. The texts were syntactically annotated according to the Universal Dependencies framework. Relative frequencies of particular syntactic relations were used to determine distances among the texts. The distances were calculated by the Cosine Delta method. Then, a hierarchical cluster analysis was performed. The results show that the method can contribute to the identification of particular presidents.

Keywords: authorship, syntax, president's speech

1. Introduction

Writing a text (or producing a speech) involves, among others, selecting the appropriate linguistic means that are used to express the intended content. Obviously, many different factors influence the selection, such as e.g. age, education, gender, the goal of writing/speaking and writer's/speaker's individuality. All of these factors leave a "footprint" in the text or speech. The main goal of stylometry to identify these "footprints". In stylometry (and in authorship analysis, its biggest domain), many approaches focusing on different language properties have been developed to date (for an overview, see Neal *et al*., 2018).

However, only a relatively small number of them concentrate on syntax (e.g.; Baayen *et al*., 1996; Sidorov *et al*., 2014; Soler-Company and Wanner, 2016). It is not surprising especially if one realizes that until relatively recently it was not easy to obtain a reliable automatic text annotation at the syntactic level. Moreover, it is an open question to which degree syntax can mirror the

authorship.
In this study, we examine whether relative frequencies of syntactic relations, as they are defined within the Universal Dependency approach (De Marneffe *et al*., 2021), reflect the authorship and whether they can be used for authorship analysis. We applied the method to language material consisting of a sample of Czech and Czechoslovak presidents delivered on the occasion of the end of the year (see Section 2).
We chose this material because the previous analysis based on the lexical characteristics of the text (cf. Kubát *et al*., 2021) brought very good results with respect to the authorship identification in this language sample. The main goal of our study is to find out whether particular authors leave beside the lexical footprint also a syntactic one.

2. Language material and methodology

For the analysis, we used a corpus consisting of 82 annual speeches of chosen Czechoslovak and Czech presidents.[1] The speeches were delivered at the end or beginning of the year. They usually summarize important events from the past year and gave an outlook for the upcoming year. The authorship analysis of presidential speeches faces several problems, which should be mentioned here. First, it is known that some presidents are real authors of their speeches (e.g. Václav Havel and Václav Klaus), while speeches of some others were written in cooperation with their collaborators. F
rom this point of view, the authorship analysis of these texts may be seen as questionable. However, each president is responsible for his speech and he hardly be expected to make a speech that he has not read beforehand and with which he disagrees (including formal characteristics of the text). Thus, in our approach, the notion of "authorship" expresses taking the political responsibility for the text.[2]
Because it was shown that there are considerable differences between the style of speeches from the communist era, on the one hand, and the democratic era, on the other, (cf. Kubát *et al*., 2021, where also the historical context is presented) we decided to split the original sample into two parts and analyze them

[1] We omitted Masaryk's speech because he delivered the only one. Hácha's speeches were not included in this research because he cannot be assigned to either of the groups consisting of communist or democratic presidents.
[2] A similar approach was used by Tuzzi et al. (2010)
[3] We also note that results are much worse (no clear pattern for speeches by particular presidents is visible) if the corpus is not split into these two parts.

separately.[3] The first sample consists of speeches given by democratic presidents (it contains 64,530 tokens), while the second sample consists of speeches given by communist presidents (it contains 70,787 tokens). Basic characteristics of the corpus are provided in Kubát *et al*. (2021).[4]

The texts were syntactically annotated according to the Universal Dependencies framework (https://universaldependencies.org). Based on this annotation, each text was transformed into the sequence of syntactic relations, such as e.g. nominal subject, object, adverbial modifier (see https://universaldependencies.org/u/dep/index.html for the complete list). For illustration, we present an example of the transformation of the first sentence taken from Havel's speech given in 1995: "Milí spoluobčané, když jsem vás před pěti lety v tento den poprvé oslovil, řekl jsem, že do našich srdcí se opět vrátila naděje."[5]

"amod nsubj mark aux obj case nummod obl case det obl obl advcl root aux mark case det obl expl:pv advmod ccomp obj"

Then, relative frequencies of syntactic relations were used to evaluate distances among particular texts. The distances were calculated by the Cosine Delta method (Smith & Aldridge 2011). The choice of the method was motivated pragmatically, it gives better results than other methods used. It is also ranked among the most popular ones (probably because it brings the best results in analyses of this kind). Finally, a hierarchical cluster analysis was performed. The analysis was executed by R package Stylo (Eder *et al*., 2016).

For calculation of distances, we use 22 most frequent syntactic relations in the case of democratic addresses and 18 most frequent syntactic relations in the case of communist addresses. T

he number of chosen syntactic relations was selected as follows. We checked results based on all possible configurations (number of the most frequent units, increment, culling, n-gram size) and chose the one that gave the best results for each sample separately (except for the number of the most frequent syntactic relations, we used the same settings for both samples: increment = 0, culling = 0, n-gram size = 1). We are aware that the approach is not general, but heuristic only. However, given that the primary goal of the study is to find out if the syntactic relations can be used for the authorship analysis, this approach is justifiable, in our opinion.

[4] For details, see https://cechradek.cz/data/cech_et_al_syntactic_footprint_list_of_texts.pdf

[5] "Dear fellow citizens, when I first addressed you on this day five years ago, I said that hope had returned to our hearts." (translated by the authors of this paper).

3. Results

Results, presented in Figures 1 and 2, show that it is possible to identify some tendencies in clustering which correspond to the authorship. One can determine four clusters among the democratic presidents (Figure 1). The most homogeneous cluster (concerning the authorship) comprises speeches given by Havel (at the bottom of the tree). However, it does not cover all of Havel's speeches (only 8 out of 14). A compelling result represents the cluster containing Beneš's speeches where we can find all his addresses, plus one Klaus gave. Similarly, we can evaluate the cluster comprising Klaus' speeches where are 9 out of 10 his total, plus one Havel's speech.

The last cluster contains all of Zeman's speeches, plus four given by Havel. To sum up, in the case of both Beneš and Zeman, all their speeches fall to particular clusters. It is also similar in the case of Klaus (with one exception). Although Havel appeared in three different clusters, we can trace a clear tendency to form a common cluster in at least some of his speeches.

The analysis of communist presidents does not bring such directly interpretable results. Nevertheless, some clear tendencies are visible. First, there are 10 out of 11 speeches in the Novotný's cluster, including one of Gottwald's addresses. Next, all of Zápotocký's speeches fall to the single cluster, but its homogeneity is disrupted by three of Husák's addresses. All other Husák's speeches (11 out of 14) create a separate cluster, filled by three of Svoboda's speeches and one of Zápotocký's. We can also find Gottwald's cluster, where are 3 out of 5 speeches, and Svoboda's one, where are 3 out of 6 speeches.

4. Discussion and conclusion

Several issues should be discussed before we present the conclusions. First, this analysis's inventory of syntactic relations is relatively small (it consists of 38 relations). Using all these relations (their relative frequencies) did not bring satisfactory results. Therefore, we determined the number of syntactic relations empirically for each sample (cf. Section 2). We found out that the results are very sensitive to the number of used relations.

Even minor changes in this parameter led to different clustering, although some major tendencies remained evident for all settings. It is why we used different numbers of relations for the two samples.

Second, proportions of syntactic relations reflect only a part of the syntax used by the author. Even if the same (or very similar) proportions appear, other syntactic characteristics can be quite different (e.g., the complexity of the syntactic tree or sentence length). If the aim is to analyze the "syntactic footprint" of the author from a more general point of view, a combination of several syntactic characteristics is necessary. On the other hand, the results presented in this study show clear tendencies in author clustering, even though they are based

only on proportions of syntactic relations. Thus, it seems reasonable to assume the existence of a specific author's "syntactic style".

Third, in this study we introduced a specific approach to the authorship ("authorship" as the political responsibility for the text) caused by the unique nature of political speeches. In particular cases, the issue of authorship is sometimes more complicated. For instance, we can see a relatively high dispersion of Svoboda's speeches. However, if we consider the president's deteriorating health (he suffered a stroke during his presidency, was later in a coma, etc.), the dispersion is not surprising. Among the democratic presidents, Havel's speeches are the most dispersed. One of possible explanations is the fact that he was a dramatist, and as such he can be supposed to display a higher variability in style.

To sum up, the results show that the method can contribute to the identification of particular presidents. Although the text clustering using the most frequent lemmas (Kubát *et al*., 2021) is more accurate, syntactic relations are nevertheless surprisingly efficient, especially if one considers the limited number of syntactic functions. It is also possible to enhance authorship attribution accuracy by combining these two approaches.

Acknowledgement

Miroslav Kubát and Radek Čech were supported by the Czech Science Foundation (GAČR), project GA22-20632S, Quantitative Syntactic Stylistics of Contemporary Written Czech. Ján Mačutek and Michaela Koščová were supported by the project VEGA 2/0096/21.

References

Baayen H., Van Halteren H. and Tweedie F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3): 121-132.

De Marneffe M. C., Manning C. D., Nivre J. and Zeman D. (2021). Universal dependencies. *Computational Linguistics*, 47(2): 255-308.

Eder M., Rybicki J. and Kestemont M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, 8(1): 107-121.

Kubát M., Mačutek J. and Čech R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1): 138-152.

Neal T., Sundararajan K., Fatima A., Yan Y., Xiang Y. and Woodard D. (2018). Surveying Stylometry Techniques and Applications. *ACM Computing Surveys*, 50(6): 1–36.

Sidorov G., Velasquez F., Stamatatos E., Gelbukh A. and Chanona-Hernández L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3): 853-860.

Smith P. and Aldridge W. (2011). Improving authorship attribution: Optimizing Burrows' delta method. *Journal of Quantitative Linguistics*, 18(1), pp. 63-88.

Soler-Company J. and Wanner L. (2016). Authorship attribution using syntactic dependencies. In Angulo, C. and Godo, L., editors, *Artificial Intelligence Research and Development*, pp. 303-308. IOS Press.
Tuzzi A., Popescu I. I. and Altmann G. (2010). *Quantitative analysis of Italian texts*. RAM-Verlag.

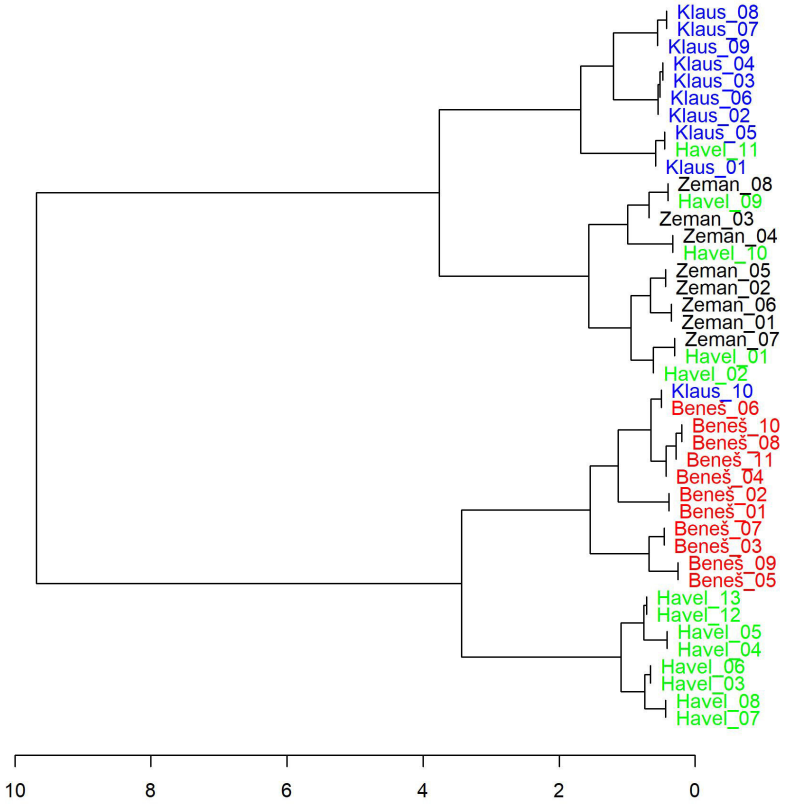


Figure 1. Cluster analysis of annual addresses given by democratic presidents.

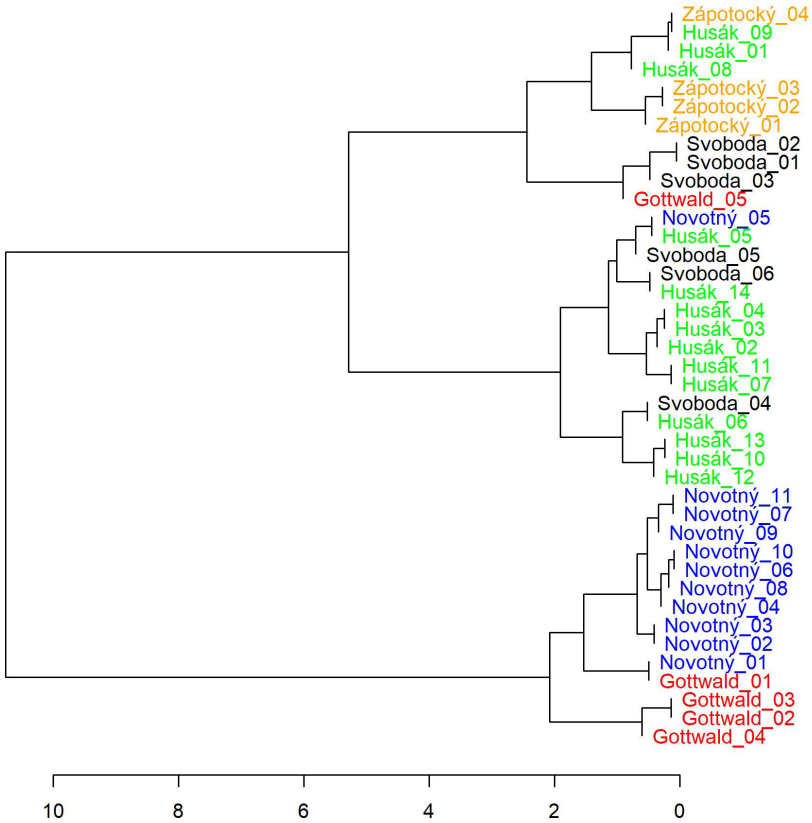


Figure 2. Cluster analysis of annual addresses given by communist presidents.

A methodological strategy for assessing the Markovian stochastic structure of the text analysis-based rank-size laws

Roy Cerqueti[1,2,3], Valerio Ficcadenti[2],
Gurjeet Dhesi[2], Marcel Ausloos[4,5]

[1]Sapienza University of Rome – roy.cerqueti@uniroma1.it
[2]London South Bank University – ficcadv2@lsbu.ac.uk
dhesig74@gmail.com
[3]University of Angers
[4]University of Leicester
[5]Bucharest University of Economic Studies
– ma683@le.ac.uk

Abstract
This paper addresses a very important problem in text analysis, namely the study of text generation. In this context, we outline the strategy to identify a stochastic process that can capture the characteristics of a system of properly ranked textual data. The problem is studied at two different levels: on the one hand, a rank size law is considered to describe the unified system generated by the words pronounced in a speech when taken with their frequency and ranked in descending order. On the other hand, a Monte Carlo Markov Chain procedure is proposed for obtaining a Markovian process having the obtained rank size law as asymptotic distribution. The methodology is described on the relevant example of the hapaxes in a corpus collecting a large amount of speeches delivered by the US Presidents.

Keywords: Text analysis, Monte Carlo Markov Chain, Zipf-Mandelbrot Law, Ranked data, Hapax Legomena.

1. Introduction

Textual data analysis is a relatively recent field of research that deals with exploring written texts with statistical and quantitative methods (see e.g., Coulthard, 2002; Kuckartz, 2014). Such a strand of studies is at the center of the current scientific debate, for two main reasons.

On one hand, text analysis requires dedicated techniques and methods, because the words in a text have multiple levels of interpretation, ranging from their structure as strings of characters to their meaning (semantics) to their use (pragmatic). Words do not really represent quantitative variables, but they can

be easily put in relationship with quantitative terms by considering their frequencies on a corpus. In addition, a text may contain uncommon occurrences like typos or even complicated mathematical symbols; such cases are quite outside the common meaning of the term "word" as a string of characters taken from an alphabet, and need a particular attention.

From this, it follows that a written text sometimes represents a complex dataset, to be managed in a non-automatic way by techniques that must always be supervised by the researcher. Therefore, text analysis is a challenging field for statisticians, mathematicians and computer scientists.

On the other hand, texts are means of communication and information, they have a global content that is based on the meanings of the single words – but, sometimes, goes beyond them. Therefore, it is essential to explore the semantic and lexical dimensions of texts in order to obtain information about their content.

This paper offers a particular perspective on text analysis, addressing an issue related to the structure of texts. Specifically, we deal with the challenging problem of text generation, with reference to words in a corpus classified on the basis of their frequencies. We follow a rank-size approach, by taking words' frequency as their size. Then, we sort the words in descending order, so that the most frequent word has rank 1. Next, we look for the descending curve that optimally approximates the ranked data through a best-fit procedure. Finally, we advance a procedure for building a stochastic process having the best-fit curve as asymptotic distribution.

The method needs an empirical example to be properly presented. The proposal advanced in this paper is to consider the corpus made of a large collection of speeches given by the US Presidents.

Data are retrieved from the Miller Center database, which hosts a large number of official speeches of the US Presidents (see http://millercenter.org). We finally have a collection of about one thousand speeches, from George Washington (Inaugural address,1789) to Donald Trump (Address to Joint Session of Congress, 2017).

Noticeably, the communication strategies of the US Presidents – or the candidates to presidency -- have attracted the attention of a rather large set of scholars (see e.g., Cochrane *et al*., 2014, Ficcadenti *et al*., 2019; Frimpon, 2013; Takikawa and Sakamoto, 2020).

We specifically refer to the frequencies in the corpus of terms pronounced only once in any individual speech, the so-called hapaxes. We stress that they are hapaxes at the level of single speech (text), not for the entire corpus. The considered empirical instance is relevant for two main reasons. First, the speeches of the US Presidents are written with extreme rigor, and the selection of the employed words is made with great care. Second, hapaxes represent a well-known measure of language richness. Moreover, hapaxes respond to non-trivial

communicative strategies, in that they often encapsulate a message that is not really emphasized but still wants to be delivered, perhaps to a specific and qualified target audience.
The considered empirical instance leads to the opportunity of selecting a Markov chain as generating stochastic process. Indeed, hapaxes can be viewed as the outcome of a memoryless process when they are taken on a sequential basis – hence, they are effectively generated by a Markovian process (see e.g. the so-called mental lexicon in Allahverdyan *et al*., 2013).
This paper moves from and extends the rank-size analysis for the hapaxes of the US Presidents speeches proposed in Ficcadenti *et al*. (2020), where the authors discuss the core of the hapaxes used by the Presidents by adapting the concept of the bibliometric H-index. We here consider the rank-size results obtained in the quoted paper and elaborate on how one can effectively derive a Markovian process having the calibrated rank-size law as asymptotic distribution.
The strategy for the assessment of the optimal Markov chain having the rank-size law as asymptotic distribution is described through a Monte Carlo Markov Chain (MCMC) procedure, in the light of the Metropolis-Hastings algorithm (see e.g., Brooks *et al*., 2011). We also propose some theoretical insights to treat properly the derivation of such a data generating process.

2. Outline of the rank-size analysis for the considered textual data

We now outline the rank-size analysis implemented in Ficcadenti *et al*. (2020), addressing the interested reader to the quoted paper for the details.
As preannounced in the Introduction, the Presidents' speeches were downloaded from the Miller Center website. After a treatment of the available data, we selected 951 speeches covering a period from 1789 (President Washington) to 2017 (President Trump). We selected the words pronounced once in all the individual speeches and stored them. The resulting list includes all the hapaxes at speech level, each of them having a frequency describing the number of speeches including them. The frequencies range from 1 to 250. Then, such words words were placed in descending order, assigning rank r=1 to the word with frequency 250. Next, a best-fit procedure was applied to test the validity of a Zipf-Mandelbrot law for approximating the scatter plot, according to the following formula:

$$f(r) = \frac{\alpha}{(\beta + r)^{\gamma}} \tag{1}$$

where r is the rank, α, β, γ are the positive parameters to be calibrated.

The best-fit exercise leads to statistically satisfactory results, with calibrated parameters given by $\widehat{\alpha} = 6.029 \times 10^8$, $\widehat{\beta} = 2540$, $\widehat{\gamma} = 1.896$.

3. Strategy for a Markovian generator of ranked hapaxes

We start from the Markovianity of the process describing the formation of the hapaxes, opportunely ranked in decreasing order. Of course, it is not reasonable to generate words; rather than this, we generate ranks. As a premise, we imagine that the speeches of the Presidents are concatenated and form an uninterrupted and unique chronological flow of words, being separated only in the phase of the identification of the hapaxes.

So, we pay attention to the following problem: suppose that a President has pronounced a hapax that in the global context is placed at rank i. What is the probability that the successive hapax will be placed at rank j?

Providing a response to the proposed question requires in essence to search for a discrete-time stochastic process X having the set of the ranks $\{1,\ldots,250\}$ as states space and satisfying the memoryless property, i.e.: a Markov chain of order 1. For such a process, the probability distribution of the realizations taken by X at a given step depends on the realization observed at the previous step.

An appropriate strategy for dealing with this problem consists of two steps.

In the first step, one needs to understand if the memoryless Markov property is empirically true or not. Only if such a condition is verified, we can reasonably proceed to the second step for a complete treatment of the problem.

In the second step, we build the Markov chain through an MCMC procedure and verify that the obtained Markov chain is a reliable representation of the original sample. The basic idea is to perform a large number of extractions from the obtained stochastic process – thus simulating a large number of asymptotic distributions – and compare these synthetic distributions with the empirical one. From a methodological point of view, the comparison between two distributions can be explored through similarity tests, such as the Kolmogorov-Smirnov. Naturally, one will derive a distribution of similarity tests related to the distance between the synthetic samples with the real one. Such a distribution should present statistical properties leading to the soundness of the asymptotic distribution of the Markov chain with the rank-size law.

The MCMC procedure can be substituted or complemented by a more theoretical perspective. In fact, finding a stationary distribution of a Markov chain is fundamentally a linear algebra problem; it can be addressed with linear systems theory. The advantage of the theoretical procedure is that it is free from the computational biases. However, the linear algebra method is obviously less efficient in the presence of very large empirical samples; moreover, it may lead to a not unique parametric solution.

4. Conclusive remarks

This paper presents a methodological strategy for detecting the stochastic structure of the rank-size laws, with particular reference to a peculiar case of textual analysis procedure. Specifically, we refer to the hapaxes in the texts collected in the corpus made by several speeches of the US Presidents.

Such an instance is of particular interest. Indeed, as mentioned above, such ranked data can be viewed as the outcomes of a memoryless process – i.e,, a Markov chain of order one.

The procedure can be applied to other rank-size contexts. More in detail, one can adopt a MCMC approach for cases associated to memoryless processes; alternatively, one can implement a preliminary step to detect the most appropriate stochastic process behind a group of ranked data; then, one can implement a tailored procedure for calibrating the related process.

Details on the development of the generation of the Markov chain having the rank-size law associated with the hapaxes as invariant distribution are currently under investigation by the authors.

References

Allahverdyan A. E., Deng W. and Wang Q. A. (2013). Explaining Zipf's law via a mental lexicon. *Physical Review E*, 88(6), 062804.

Brooks S., Gelman A., Jones G. and Meng X. L. (Eds.). (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Cochran J. J., Curry D. J., Radhakrishnan R. and Pinnell J. (2014). Political engineering: optimizing a US Presidential candidates platform. *Annals of Operations Research*, 215(1), 63-87.

Coulthard M. (Ed.). (2002). *Advances in written text analysis*. Routledge.

Ficcadenti V., Cerqueti R. and Ausloos M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications*, 123, 127-142.

Ficcadenti V., Cerqueti R., Ausloos M. and Dhesi G. (2020). Words ranking and Hirsch index for identifying the core of the hapaxes in political texts. *Journal of Informetrics*, 14(3), 101054.

Frimpon M. F. (2013). A multi-criteria decision analytic model to determine the best candidate for executive leadership. *Journal of Politics and Law*, 6(1), 111-127.

Kuckartz U. (2014). *Qualitative text analysis: A guide to methods, practice and using software*. Sage.

Takikawa H. and Sakamoto T. (2020). The moral–emotional foundations of political discourse: a comparative analysis of the speech records of the US and the Japanese legislatures. *Quality & Quantity*, 54(2), 547-566.

Analyse du profil sémantique des expressions biodiversité et diversité biologique dans la presse écrite française

Marie Chandelier, Sascha Diwersy

Université Paul Valéry Montpellier – marie.chandelier@univ-montp3.fr
sascha.diwersy@univ-montp3.fr

Abstract
Environmental crisis has led to the emergence of new categories to define nature. In particular, the expressions *biological diversity* and *biodiversity* have been created in the scientific field and disseminated to the public audience during the Rio Earth Summit in 1992. Neologism *biodiversity* has then been widely used in the political, institutional, activist and media spheres. Despite its large appropriation, the term is also being challenged by a critical metalinguistic discourse centered on the neologism itself and on its semantic instability. In order to observe the evolution of semantic properties associated with *biological diversity* and *biodiversity*, we analysed their combinatorial profile in the daily newspaper *Le Monde*, from 1979 to 2018 (3112 articles). We focused on *biodiversity* collocates to study changes of meaning along the period. Our research showed a variation of word associations between 1990-2004 and 2005-2018. While *appauvrissement* (depletion), *menace* (threat) and *maintien* (maintaining) are associated with the first period, *effondrement* (collapse), *érosion* (erosion) and *reconquête* (regaining) are specific to the second period. This variation reveals an evolution of Le Monde framing of biodiversity crisis from a current threat to a destruction phenomenon that is already happening.

Keywords: corpus linguistics, collocation, discourse analysis, neologism, biodiversity

Résumé
La crise environnementale est marquée par l'émergence de nouvelles catégories pour référer à la nature. C'est le cas des expressions diversité biologique et biodiversité, élaborées dans le champ scientifique et diffusées dans l'espace public lors de la médiatisation du Sommet de la Terre à Rio en 1992. Le nom biodiversité fait alors l'objet d'une appropriation rapide dans les sphères politique, institutionnelle, militante et médiatique. L'usage croissant de ce terme s'accompagne néanmoins d'un discours métalinguistique critique, centré sur la forme même du néologisme et sur le caractère peu stabilisé de sa définition. Notre étude, réalisée sur un corpus d'articles publiés dans le quotidien Le Monde entre 1979 et 2018 (3112 articles), se focalise sur l'analyse du profil combinatoire des expressions biodiversité et diversité biologique, afin d'identifier les propriétés sémantiques partagées par ces deux expressions et celles qui sont propres

à chacune. Du point de vue de l'évolution diachronique, nous nous sommes centrés sur l'évolution des collocatifs du néologisme biodiversité. Les premières périodes (1990-1994 ; 1995-2004) sont caractérisées par un usage significatif des termes appauvrissement, menace et maintien, sous-employés au cours de la dernière période (2005-2018). En revanche, entre 2005 et 2018, ce sont les termes effondrement, érosion et reconquête qui sont significativement plus utilisés, traduisant ainsi l'évolution vers la représentation d'une destruction avérée du vivant et de la nécessité, non plus de maintenir la biodiversité, mais de la regagner. Notre étude montre que par ses contextes d'emploi dans le quotidien Le Monde, le mot biodiversité intègre à sa définition le péril que les activités humaines engendrent pour les espèces et les écosystèmes, et le rôle majeur de l'homme pour permettre leur sauvegarde.

Mots clés: linguistique de corpus, profil combinatoire, analyse du discours, néologie, biodiversité

1. Introduction

En invitant à repenser les relations entre l'humain et la nature, la crise environnementale a favorisé l'émergence d'expressions et de mots nouveaux tels que *environnement*, *développement durable* ou encore *biodiversité*. Voués à redessiner les contours de référents socio-écologiques en mutation, les termes constituent, comme le note la linguistique Sandrine Reboul-Touré, des mots-témoins pour l'analyse du discours (Reboul-Touré 2000). Le contexte de crise dans lequel ils apparaissent les rapproche du statut de formule (Krieg-Planque 2009). En effet, ces unités lexicales acquièrent, par leurs conditions d'émergence, un caractère de référent social, alors que la multiplicité des usages dans l'espace public favorise une instabilité sémantique. Celle-ci se traduit par une diversification des définitions associées aux termes et par la présence d'un discours métalinguistique critique. Dans une précédente étude, nous nous sommes intéressés aux énoncés définitionnels associés aux expressions *diversité biologique* et *biodiversité*, qui émergent lors de la médiatisation de la crise environnementale (Chandelier *et al*. 2020). En nous appuyant sur un corpus d'articles du quotidien *Le Monde* publiés entre 1979 et 2018, nous avons montré que le terme *biodiversité* est doté de propriétés nouvelles au cours du temps, notamment métaphoriques, alors que l'expression *diversité biologique* reste essentiellement affectée au champ institutionnel. Afin d'analyser la construction du sens de ces deux expressions dans *Le Monde*, nous nous concentrerons ici sur l'analyse du profil combinatoire de *diversité biologique* et de *biodiversité*. Nous nous focaliserons ensuite sur l'évolution diachronique des collocatifs du néologisme *biodiversité*, dont les usages sont majoritaires dans notre corpus.

2. Description du corpus et méthode d'analyse

Notre étude porte sur un corpus d'articles de presse écrite publiés par le quotidien Le Monde entre 1979 et 2018 (3228 articles, 2 749 836 mots)[1]. Les documents ont été extraits à partir de la plateforme Europresse, selon la méthode du Relevance sampling (Krippendorf 2004 : 118-120). Nous avons sélectionné les articles comportant au moins une fois le mot biodiversité et l'expression diversité biologique. Le corpus ainsi constitué comprend 123 articles avec les deux expressions, 116 avec l'expression diversité biologique et 2989 avec le néologisme biodiversité. Le corpus a été annoté au niveau des lemmes, de la morphosyntaxe et des relations de dépendance syntaxique à l'aide de l'analyseur Talismane[2] (Urieli 2013). Il a été indexé avec les outils de compilation du moteur de recherche CQP (Corpus Query Processor, Evert & Hardie 2011). L'analyse de la combinatoire de biodiversité et de diversité biologique est basée en grande partie sur le traitement statistique de leurs cooccurrences lexico-syntaxiques, qui ont été extraites de façon automatique à partir des annotations en relations de dépendance dont est doté notre corpus. D'abord, nous avons établi les inventaires cooccurrentiels de biodiversité et diversité biologique en créant des listes répertoriant leurs collocatifs sous forme de triplets définis par les propriétés [lemme + partie du discours + relation de dépendance syntaxique]. Nous avons constitué un inventaire sans différentiation chronologique pour analyser les collocatifs de diversité biologique et de biodiversité, ainsi qu'un inventaire avec différenciation chronologique pour étudier l'évolution des collocatifs de biodiversité. Les inventaires chronologiquement différentiés sont basés sur une partition du corpus en périodes de cinq ans commençant par l'intervalle [1990-1994] et se terminant à l'intervalle [2015-2018]. Pour les collocatifs récupérés, nous avons ensuite effectué, selon la méthodologie relevant de l'analyse collostructionnelle (Stefanowitsch & Gries 2003), un calcul de scores d'association en termes du test exact de Fisher-Yates. Les collocatifs adjectivaux, nominaux et verbaux ont finalement été enregistrés dans des lexicogrammes (ici illustrés par les extraits présentés dans les tableaux 1 et 2) qui constituent le jeu

[1] Nous avons focalisé notre étude sur l'analyse diachronique des articles publiés dans *Le Monde*, en raison du statut historique de journal de référence attribué au quotidien et de l'accessibilité des archives entre 1979 et 2018 pour ce seul journal (https://larevuedesmedias.ina.fr/series/le-monde-itineraire-dun-journal-de-reference).

[2] Nous avons utilisé la version 5.2.0 du logiciel (téléchargeable à l'adresse https://github.com/jo-liciel-informatique/talismane/releases/tag/v5.2.0) avec une configuration de la chaîne de traitement faisant appel au modèle d'annotation Universal Dependencies (*fr-ud*). Les jeux d'étiquettes UD respectifs (parties du discours, traits morphologiques, fonctions syntaxiques) pour le français sont documentés à la page https://universaldependencies.org/fr/index.html (date de consultation : 02/01/2020).

de données principal. Nous avons appliqué à ce jeu de données plusieurs procédés lexicométriques adaptés au traitement de données diachroniques. En nous concentrant sur l'évolution diachronique des collocatifs de *biodiversité*, nous avons soumis le corpus à une périodisation automatique par classification ascendante hiérarchique par contiguïtés (CAHC), selon la méthode *Variability-based Neighbour clustering* (VNC) proposée par Gries & Hilpert (2008). Nous avons alors appliqué le calcul d'accroissements spécifiques d'une période à l'autre (Lebart & Salem 1994, 226-229), pour décrire les changements lexicaux parmi les cooccurrents de *biodiversité*.

Tableau 1. Extrait du lexicogramme sans distinction hiérarchique

Mot-pôle	**Collocatif**	**Co-fréquence**	**Score FisherYates**
biodiversité	érosion__NOUN__H[3]	149	224,9041
…	…	…	…
biodiversité	préservation__NOUN__H	97	122,7696
…	…	…	…
diversité biologique	convention__NOUN__H	36	69,6793
…	…	…	…
diversité biologique	préservation__NOUN__H	6	8,3937
…	…	…	…

Tableau 2. Extrait du lexicogramme avec distinction hiérarchique

Mot-pôle	**Collocatif**	**Co-fréquence**	**Score FisherYates**
biodiversité	érosion__NOUN__H[3]	149	224,9041
…	…	…	…
biodiversité	préservation__NOUN__H	97	122,7696
…	…	…	…
diversité biologique	convention__NOUN__H	36	69,6793
…	…	…	…
diversité biologique	préservation__NOUN__H	6	8,3937
…	…	…	…

[3] La notation « érosion__NOUN__H » indique que le lemme *érosion* avec la catégorie *nom* (NOUN) est en position de *mot recteur* (H) par rapport au mot-pôle, en l'occurrence *biodiversité*.

3. Profils combinatoires de biodiversité et diversité biologique

Le tableau ci-dessous présente les collocatifs de *biodiversité* et de *diversité biologique* selon trois types d'association[4]:

- Les collocatifs partagés par *biodiversité* et *diversité biologique* ;
- Les collocatifs propres à *biodiversité* ;
- Les collocatifs propres à *diversité biologique*.

Tableau 3. Collocatifs partagés et collocatifs propres à chaque expression

Collocatifs partagés	Collocatifs propres à *biodiversité*	Collocatifs propres à *diversité biologique*
Substantifs relevant de l'isotopie institutionnelle : *conférence* ; *convention* Substantifs relevant de l'isotopie de la crise environnementale : *érosion* ; *perte* ; *préservation* ; *protection*	Adjectifs opérant une restriction de catégorie : *animal* ; *marin* ; *terrestre*, *végétal* ; *ordinaire* ; *mondial* Adjectifs évaluatifs : *riche* ; *exceptionnel* ; *incroyable* ; *essentiel* Substantifs relevant de l'isotopie de la crise environnementale : *impact* ; *atteinte* ; *crise* ; *déclin* ; *dégradation* ; *destruction* ; *effondrement* ; *réduction* ; *menace* ; *peser* ; *défense* ; *reconquête* ; *richesse* ; *sanctuaire* ; *respect* ; *chevet* Substantifs et verbes relevant de l'isotopie de la gestion : *gestion* ; *loi* ; *plan* ; *agence*, *consacrer* ; *favoriser* ; *abriter* ; *journée* Substantifs relevant de l'isotopie scientifique : *plate-forme* ; *observatoire*, *spécialiste* ; *indicateur* ; *inventaire* ; *climat*	-

Les collocatifs partagés par *biodiversité* et *diversité biologique* relèvent du vocabulaire institutionnel (*conférence* ; *convention*) et du lexique de la crise environnementale (*érosion* ; *perte* ; *préservation* ; *protection* ; etc.). D'un point de vue syntaxique, les expressions *diversité biologique* et *biodiversité* sont employées dans des syntagmes nominaux en fonction complément. Entre 1990 et 2018, lors de la médiatisation des Conférences des Parties (COP), les

[4] Les collocatifs listés correspondent aux premiers 10% en termes de score d'association.

journalistes emploient indifféremment les syntagmes nominaux *conférence sur la diversité biologique* ou *conférence sur la biodiversité* pour désigner les Conventions sur la Diversité Biologique, illustrant ainsi la relation synonymique entre les deux expressions. En revanche, cette relation synonymique ne semble pas être maintenue hors des contextes institutionnels. À la différence des collocatifs partagés par les deux expressions, ceux qui sont propres à *biodiversité* appartiennent à des catégories grammaticales et à des isotopies variées. La prédominance de noms et d'adjectifs subjectifs (tableau 3) confère une plus forte intensité au discours relatif à la crise de la biodiversité. En particulier, l'association des termes *sanctuaire*, *respect* et *chevet* témoignent d'une appréhension symbolique de la biodiversité, comme le montre l'extrait de l'article intitulé « Biodiversité : fortes paroles, faibles moyens », publié le 7 juillet 2018 : « En elle-même, la présence d'une douzaine de ministres au chevet de la biodiversité est un signal fort. Elle montre que plus personne jusqu'au sommet de l'Etat ne peut continuer à ignorer, ou feindre d'ignorer, que la nature se meurt en silence, sous nos yeux ». Intégré au compte rendu de la visite d'un comité interministériel au Museum national d'Histoire naturelle de Paris, la locution *au chevet de*, qui signifie « auprès d'une personne alitée (généralement pour cause de maladie mortelle) » (TLFi), personnifie la biodiversité en projetant dans le même temps sa disparition prochaine. La personnification est appliquée à la nature dans son ensemble, envisagée comme un seul et même être vivant condamné à s'éteindre. L'association préférentielle du terme biodiversité à un lexique subjectif et évaluatif confirme ce qui a été observé d'un point de vue qualitatif pour l'étude des définitions et des métaphores de la *biodiversité* dans la presse écrite française et anglophone (Chandelier *et al*. 2020 ; Hellsten 2002). À la différence de l'expression *diversité biologique*, le néologisme biodiversité acquiert dans le quotidien *Le Monde* des propriétés affectives, qui le constituent en tant que référent sensible. La diversification du réseau sémantique de *biodiversité* est facilitée par l'appropriation rapide du terme et par son usage croissant dans la presse écrite depuis son émergence. Cette diversification peut être observée par l'étude de l'évolution des collocatifs de *biodiversité*, qui constitue un indicateur des changements de propriétés associées au référent.

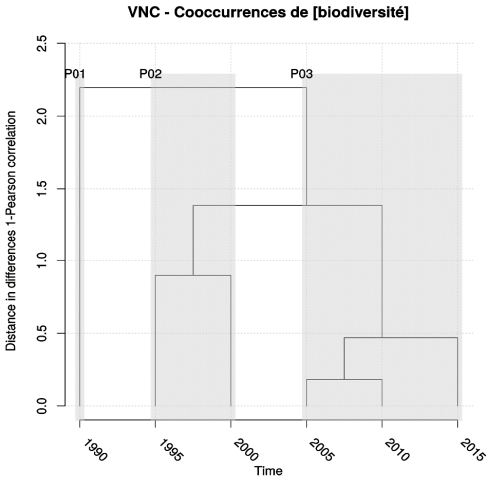
4. Analyse de l'évolution diachronique des collocatifs de biodiversité

Les collocatifs propres à biodiversité font apparaître la variété des contextes d'emploi du néologisme, du champ scientifique et institutionnel au champ militant. Afin d'identifier les changements lexicaux parmi les collocatifs de *biodiversité*, nous avons appliqué à l'inventaire cooccurrentiel du mot une périodisation automatique et analysé les changements de vocabulaire entre les périodes par le calcul d'accroissements spécifiques (Lebart & Salem 1994, 226-229).

4.1 Périodisation automatique

La périodisation automatique par classification ascendante hiérarchique par contiguïté (Gries & Hilpert 2008) appliquée à l'inventaire cooccurrentiel de *biodiversité* structure le corpus en trois parties (figure 1).

La périodisation automatique regroupe les années 1990-1994 (P01) ; 1995-2004 (P02) et 2005-2018 (P03). À partir de ce découpage, nous avons réalisé une analyse diachronique du vocabulaire associé à *biodiversité* dans le quotidien *Le Monde*.



4.2 Calcul d'accroissements spécifiques par période

À l'aide des calculs d'accroissements spécifiques (Lebart & Salem 1994, 226-229), nous avons identifié les termes sur-employés et sous-employés entre les périodes P02 (1995-2004) et P01 (1990-1994) ainsi qu'entre les périodes P03 (2005-2018) et P01 + P02 [1990-2004].

P02 vs P01		P03 vs P01+P02	
Sur-emplois	Sous-emplois	Sur-emplois	Sous-emplois
-	*convention* ; *enjeu*	*effondrement* ; *érosion* ; *reconquête* ; *exceptionnel* ; *mondial* ; *loi* ; *plan*	*fruit* ; *ressource* ; *appauvrissement* ; *menace* ; *maintien* ; *maximal* ; *global* ; *milieu* ; *convention* ; *rapport* ; *journée* ; *diversité*

Tableau 4. Calcul d'accroissements spécifiques par période

L'évolution des cooccurrents entre les différentes périodes révèle une variation de la représentation de la crise de la biodiversité (Tableau 4). La comparaison de la période P02 [1994-2005] à la période P01 [1990-1994] fait apparaître le rôle de la médiatisation du Sommet de la Terre à Rio en 1992 pour la diffusion du néologisme *biodiversité*. En effet, l'usage des collocatifs *convention* et *enjeu* demeure spécifique à la période 1990-1994, au cours de laquelle sont relayés les débats concernant la Convention sur la Diversité Biologique et ses enjeux pour la conservation du vivant. La comparaison de la période P03 aux périodes P02 et P01 traduit une évolution de la représentation de la crise environnementale, d'une dégradation en cours à une perte avérée de la biodiversité. Durant la période 1990-2004, le sur-emploi des termes *appauvrissement* et *menace* confère à la destruction du vivant un caractère processuel et imminent. En revanche, le collocatif *effondrement*, significativement plus utilisé entre 2005 et 2018 confère un stade plus avancé au déclin de la biodiversité. Du point de vue du lexique relatif à la protection de la biodiversité, la période 1990-2004 se caractérise par un sur-emploi du nom *maintien*, qui suppose que l'état de la biodiversité à l'instant considéré est encore suffisant. Le substantif *reconquête*, spécifique de la période 2005-2015, entérine quant à lui la reconnaissance de la perte de la biodiversité, et la nécessité d'une mise en œuvre d'actions destinées à la restaurer. Les occurrences du syntagme *reconquête pour la biodiversité* dans notre corpus sont essentiellement liées à la question de la « loi pour la reconquête de la biodiversité » (31 des 36 occurrences). Sa dépendance syntaxique au mot loi l'inscrit exclusivement dans le cadre politique.

5. Conclusion

L'analyse des profils combinatoires de *diversité biologique* et de *biodiversité* a permis d'identifier les contextes d'usage partagés par les deux expressions et ceux qui étaient propres à chacune dans le quotidien *Le Monde*. L'expression *diversité biologique* et le néologisme biodiversité sont tous deux employés dans des contextes relatifs à la gestion institutionnelle de la crise environnementale. À la différence du syntagme *diversité biologique*, le néologisme *biodiversité* est associé à un lexique plus subjectif et métaphorique, ce qui contribue à lui attribuer non plus uniquement des propriétés scientifiques, mais aussi celles d'un être sensible dont l'existence est mise en péril par les activités humaines. D'un point de vue diachronique, l'analyse des collocatifs de *biodiversité* dans le quotidien national met en évidence une évolution vers une description de la crise environnementale non plus seulement en tant que processus mais en tant que phénomène avéré.

Bibliographie

Chandelier M, Diwersy, S. et Paissa P. (2020). « Emergence et diffusion des formules

diversité biologique et biodiversité dans Le Monde (1979-2018) ». *7e Congrès Mondial de Linguistique Française* 78, 1-13. https://doi.org/10.1051/shsconf/20207801009
Evert S. et Hardie A. (2011). « Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium ». *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK.
Gries S. Th. et Hilpert M. (2008). « The identification of stages in diachronic data: variability-based neighbour clustering ». *Corpora*, 3, 59–81.
Hellsten I. (2002). *The politics of metaphor: Biotechnology and biodiversity in the media*. Tampere : Tampere University Press.
Krieg-Planque A. (2009). *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique.* Besançon : Presses Universitaires de Franche-Comté.
Krippendorff K. (2004). *Content analysis: an introduction to its methodology*. Londres : Thousand Oaks.
Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris: Dunod.
Reboul-Touré S. (2020). « La biodiversité : un mot-témoin pour l'analyse du discours ». *Les Carnets du Cediscor* [En ligne], 15. https://doi.org/10.4000/cediscor.2662
Salem A. (1988). « Approches du temps lexical ». *Mots*, 17, 105-143.
Stefanowitsch A. et Gries S. Th. (2003). « Collostructions: Investigating the interaction between words and constructions ». *International Journal of Corpus Linguistics*, 8, 2, 209-243.
Urieli A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de Doctorat, Université de Toulouse II le Mirail.

Rappresentazioni sociali della migrazione e dell'identità culturale in testi letterari di scrittrici e scrittori di origine africana

Giacomo Chiara, Diego Romaioli, Alberta Contarello

Università di Padova – giacomo.chiara@phd.unipd.it
diego.romaioli@unipd.it alberta.contarello@unipd.it

Abstract
In this paper we present a study on social representations of migration and cultural identity. Using a quali-quantitative method, such as the Reinert method, we have analyzed 82 literary texts written by authors of African descent that fall within the strand of Italian postcolonial literature. The results suggest that the shared knowledge of migration and cultural identity conveyed by these literary texts is not unitary but contains many elements of complexity. The migrant is represented as an illegal immigrant, a laborer, as an individual whose identity as a black Italian is not recognized, but also through a dimension of everyday life, with narratives about cooking and expression of emotions. Moreover, the figure of the migrant emerges as a writer who through writing wants to leave a sign, of testimony and/or resistance on the one hand, showing a normalization of everyday experience on the other. The work closes with some reflections on the potential of this type of material and of quali-quantitative methodologies for psychosocial research.

Keywords: social representations; migration; social psychology and postcolonial Italian literature; qualitative-quantitative methods.

Riassunto
In questo scritto presentiamo una ricerca sulle rappresentazioni sociali della migrazione e dell'identità culturale. Nello specifico, attraverso l'uso di un metodo quali-quantitativo, come il metodo Reinert, abbiamo analizzato 82 testi letterari di scrittrici e scrittori di origine africana, che rientrano nel filone della letteratura italiana postcoloniale. I risultati suggeriscono che la conoscenza condivisa della migrazione e dell'identità culturale veicolata da questi testi letterari non si presenta come unitaria bensì contiene molti elementi di complessità. Il migrante viene rappresentato come un clandestino, un bracciante, come individuo a cui non è riconosciuta la propria identità di italiano nero, ma anche attraverso una dimensione della vita quotidiana, con narrazioni sulla cucina e l'espressione di emozioni. Inoltre, emerge la figura del migrante come scrittore/scrittrice che attraverso la scrittura vuole lasciare un segno, di testimonianza e/o di resistenza da una parte, di normalizzazione dell'esperienza quotidiana dall'altra. Il lavoro si conclude con alcune riflessioni sulle potenzialità di questo tipo di materiale e delle metodologie

quali-quantitative per la ricerca psicosociale.

Parole chiave: rappresentazioni sociali; migrazione; psicologia sociale e letteratura italiana postcoloniale; metodi quali-quantitativi.

1. Introduzione

Da tempo il fenomeno della migrazione costituisce un campo di ricerca privilegiato per la psicologia e la ricerca sociale. Tuttavia, poca attenzione è stata data al rapporto tra opere letterarie e migrazione, nonostante alcuni studiosi abbiano enfatizzato come la conoscenza del mondo sociale possa essere incrementata proprio dall'analisi dei testi letterari (Contarello, 2008; Moscovici, 1986). Adottando uno sguardo psicosociale, le opere letterarie sulla migrazione divengono artefatti sociali attraverso i quali è possibile accedere alle esperienze migratorie in prima persona; esperienze che una volta narrate e divulgate possono contribuire in modo sostanziale allo sviluppo e alla modificazione di rappresentazioni sociali (Contarello *et al*., 2011). Con rappresentazioni sociali ci si riferisce a un sistema di credenze, idee, valori e pratiche sociali attraverso cui la conoscenza si organizza e orienta gli individui nel mondo sociale fornendo codici comunicativi per lo scambio sociale (Moscovici, 1976). In questa prospettiva. l'interesse verte su quelle forme di pensiero pratico, sociale e condiviso che si costruiscono e si modificano in relazione a cambiamenti psicosociali e societari (Moscovici, 1976; Jodelet, 1989; Mancini *et al*., 2007). Diversi studi hanno rivolto particolare attenzione ai processi di comunicazione e quindi alle fonti maggiormente implicate nella costruzione di come l'immigrazione viene pensata, discussa e vissuta, costituite dalla stampa (Montali *et al*., 2013; Mazzara *et al*., 2020) e dai social media (Mannarini *et al*., 2020; de Rosa *et al*., 2021). Gli studi sulle rappresentazioni sociali della migrazione nei media italiani ed europei hanno evidenziato il predominio di una visione negativa della migrazione (Maneri, 2011; Triandafyllidou, 2013), vista come minaccia all'autenticità culturale (Triandafyllidou, 1999) o nei termini di una questione di sicurezza da affrontare e come minaccia interna (Mazzara *et al*., 2020).La letteratura sulle rappresentazioni sociali dei migranti forzati trasmette rappresentazioni basate sulla vittimizzazione da un lato e sulla minaccia che possono rappresentare per il paese ospitante dall'altro (Bottura & Mancini, 2016). Nel contesto italiano il migrante è rappresentato come un nemico: come un invasore, un alieno o un criminale (Binotto, 2015). Tuttavia, come scrive Jacomella (2015), il punto di vista dei migranti nei media italiani è sottorappresentato. Il nostro interesse è quello di indagare le rappresentazioni sociali della migrazione e dell'identità culturale prendendo in considerazione testi letterari di scrittrici e scrittori africani o italiani con origine africana, che rientrano nella cosiddetta letteratura italiana postcoloniale (Romeo, 2018). Dal nostro punto di vista riteniamo utile fare ricerca psicosociale con i testi letterari

attraverso un'analisi automatica del contenuto (Tuzzi, 2003), e nello specifico analisi quantitative di dati testuali (QATD – Quantitative Analysis of Textual Data), ispirandoci alla topic detection (Sbalchiero, 2018) e ai metodi quali-quantitativi come, ad esempio, il metodo Reinert (Reinert, 1983, 1990). La nostra posizione come ricercatori in questo caso è in linea con un approccio *distant reading* (Moretti, 2013), definito come un modo per contare, mappare o visualizzare i testi. In altre parole, si tratta di una lettura a distanza, nella quale, attraverso l'ausilio di strumenti computazionali, si ottiene il corpus testuale in una forma sintetica, immediata e capace di rispondere a specifici quesiti.

1.1. Domande di ricerca

Le nostre domande di ricerca sono: Quali rappresentazioni sociali della migrazione nel contesto italiano vengono veicolate dai testi letterari postcoloniali di scrittrici e scrittori africani? Come viene costruita l'identità culturale in questi testi?

2. Metodo

2.1. Corpus testuale e criteri di selezione

Il corpus testuale è composto in totale da 82 testi letterari che rientrano nelle tre fasi principali della letteratura italiana postcoloniale: nello specifico, 45 romanzi e 37 brevi racconti pubblicati dal *Premio letterario per immigrati Eks&Tra*. I testi sono stati selezionati attraverso la banca dati «Basili&Limm», seguendo la voce "testi autobiografici", ai quali sono stati aggiunti altri testi di recente pubblicazione.

2.2. Procedura e Analisi

Dopo una prima fase di trattamento del testo (Bolasco, 1999), per disambiguare alcune parole, sul materiale testuale è stata effettuata un'analisi del contenuto, nello specifico un'analisi gerarchica discendente (Reinert, 1990) con il software IRaMuTeQ (Ratinaud & Merchand, 2012). Questo metodo è stato sviluppato nel campo delle scienze psicosociali allo scopo di gestire grandi corpus testuali e di colmare il divario tra l'analisi qualitativa e quantitativa del testo (Sbalchiero, 2018). Nello specifico, il metodo Reinert è un metodo procedurale per l'analisi dei dati testuali in grado di analizzare la distribuzione dei sintagmi all'interno di un testo. La classificazione discendente gerarchica è un metodo di classificazione che suddivide gli enunciati (U.C.E.) che compongono un testo in classi di significato, definiti "mondi lessicali", sulla base delle co-occorenze delle parole in porzioni definite di testo rilevate attraverso la misura del X^2 (Ratinaud & Marchand, 2012; Sbalchiero, 2018). Il dendrogramma che il software genera, di seguito presentato in Figura 1, contiene quelle con valore di probabilità $p < .05$. Attraverso questo metodo è possibile ottenere una gerarchia di classi e una

rappresentazione grafica della divisione delle classi, la quale illustra in che modo le varie classi sono connesse fra loro.

3. Risultati

L'analisi svolta ha permesso di rilevare 25547 segmenti testuali, 82734 forme, 2241257 occorrenze, 21252 hapax (0,95% di occorrenze, 44,24% di forme), ha classificato 19916 segmenti su 25547 (77.96%).

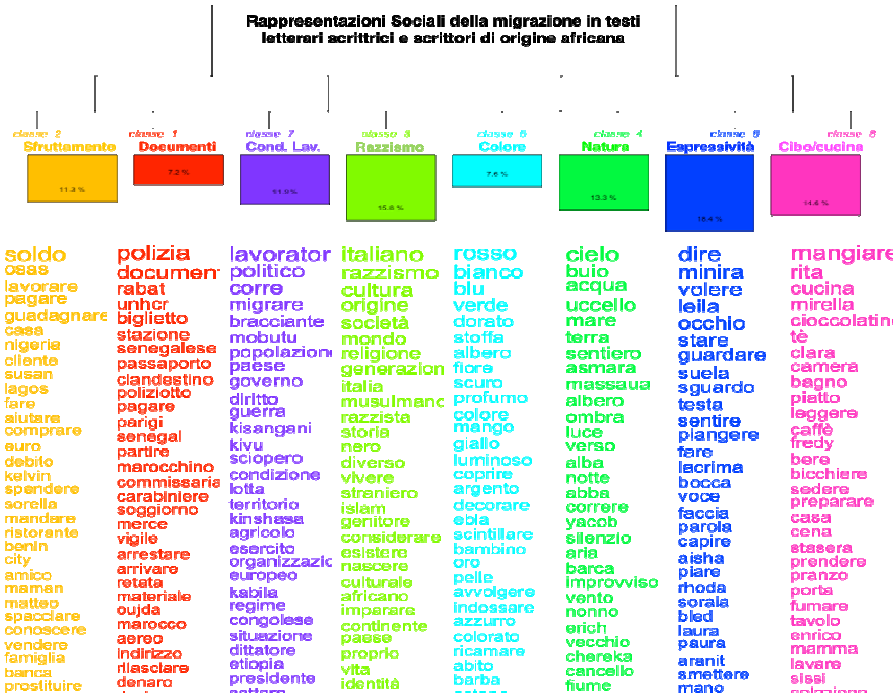


Figura 1. Dendrogramma di classificazione, dimensione delle classi e lista delle parole più caratteristiche in ogni classe (in ordine crescente di collegamento chi2 alle classi)

Figura 1 offre una panoramica del profilo di queste classi. L'analisi gerarchica discendente indica otto mondi lessicali suddivisi in due macro-gruppi. Il macro-gruppo di destra contiene mondi lessicali che si riferiscono alla dimensione della vita quotidiana, ovvero alla cucina e al cibo (classe 8), al tema della natura e dell'ambiente esterno (classe 4), del colore (classe 5) e dell'espressività delle emozioni (6). Il macro-gruppo di sinistra illustra le difficoltà della vita del migrante nel contesto italiano, a partire dai documenti (classe 1), le difficoltà lavorative e condizioni di sfruttamento lavorativo, eventuali debiti da pagare a sfruttatori e/o trafficanti (classe 2), fino alle situazioni di discriminazione e di

razzismo, così come al disagio di vedere misconosciuta la propria identità di italiano nero (classe 3). Per ragioni di spazio, ci concentriamo sul macro-gruppo di sinistra, rilevante ai fini delle nostre domande di ricerca (da sinistra a destra in Figura 1).
La classe 2 rappresenta l'11.3 % del corpus testuale e contiene termini che evocano la condizione di sfruttamento lavorativo e/o sessuale delle migranti. Sono presenti termini come *prostituire, maman, debito, soldo, Nigeria, cliente, lagos*, ma anche *spacciare, lavorare, guadagnare*. Di seguito viene riportato un estratto:

anche loro fanno i riti voodoo prima di partire e hanno paura delle conseguenze e così pure gli uomini diventano allo_stesso_tempo sfruttati e sfruttatori fingono di essere innamorati delle ragazze per spingerle sulla strada e chiedere soldi poi magari si sposano in nigeria e mandano a casa i soldi che le loro ragazze guadagnano in italia con la prostituzione è un altro crudele raggiro per queste donne alcuni uomini invece tornano in nigeria per sposarsi e_poi portano le mogli qui e le fanno prostituire in nigeria non sarebbe nemmeno concepibile una cosa del_genere
*[**** *ID_57 *fase_3 *genere_donna *areageo_africa *scrittura_collaborativa *anno_2018 *autore_Okoedion&Pozzi]*

La classe 1 rappresenta il 7.2% del corpus testuale e include contenuti sui *documenti*, sul *permesso di soggiorno* e sulla vita da *clandestino*. Sono anche presenti le situazioni e figure istituzionali a cui il migrante clandestino può incorrere, come interazioni con la *polizia*, il *poliziotto*, il *carabiniere*, e luoghi come la *stazione* e il *commissariato*:

se fossi stato accolto nella struttura avrebbero dovuto compilare una scheda con i miei dati e avrebbero subito scoperto che non avevo il permesso di soggiorno e in quanto clandestino sarei stato rimpatriato poco dopo dalla polizia non_so come ma per qualche strano miracolo tutto questo non è successo sono stato accolto dai medici al pronto soccorso senza alcuna formalità mi hanno solo chiesto se avessi un documento ho dato il mio passaporto che avevo sempre con me ed è filato tutto liscio
*[**** *ID_77 *fase_3 *genere_uomo *areageo_africa *scrittura_collaborativa *anno_2016 *autore_Kane&Garaw]*

La classe 7 rappresenta l'11.9% del corpus testuale e riassume la *condizione* del *lavoratore* migrante, che è quella del *bracciante* senza *diritti* e che *sciopera*, ma sono presenti anche contenuti sulla *guerra*, sulle *dittature*:

se vogliamo ridare dignità e far riconquistare i diritti sindacali e sociali attraverso un processo di organizzazione e sindacalizzazione lungo tutta la filiera agricola dai campi fino alla tavola dobbiamo partire dal protagonismo delle braccianti e dei braccianti in una prospettiva di alleanza con i contadini e gli agricoltori da un lato e i consumatori

dall altro per chiedere uguale lavoro uguale salario un cibo sano e una vita degna
*[**** *ID_79 *fase_3 *genere_uomo *areageo_africa *scrittura_singola *anno_2020 *autore_Soumahoro]*

La classe 3 rappresenta il 15,8% del corpus testuale, seconda per ordine di importanza, e include contenuti che enfatizzano il tema dell'*identità* e del *razzismo*, della *religione* e dell'essere *diverso*. Essere *italiano* con la pelle *nera* significa essere considerati *immigrato* e *straniero*. Sono presenti anche contenuti sulla nuova generazione. Infatti, sono scrittrici e scrittori italiani neri nati in Italia da genitori migranti a scrivere di identità, la quale diventa *un'arena di scontro politico* e *corpi spersonalizzati senza identità*. Come scrive Scego: *Credo di essere una donna senza identità. O meglio con più identità*.

è come essere trattati da cittadina di serie b sì sei come gli altri ma non proprio uguale la mia identità è un arena di scontro politico in cui è un altra persona a decidere se io sia italiana o meno
*[**** *ID_69 *fase_3 *genere_donna *areageo_africa *scrittura_singola *anno_2020 *autore_Obasuyi]*

le persone nere sono appunto corpi spersonalizzati senza identità pensieri opinioni le persone nere sono a tratti degli invasori oppure dei cuccioli da salvare sono da sfruttare oppure da nominare per appuntarsi la propria medaglietta di antirazzista perfetto
*[**** *ID_69 *fase_3 *genere_donna *areageo_africa *scrittura_singola *anno_2020 *autore_Obasuyi]*

Credo di essere una donna senza identità. O meglio con più identità. Chissà come saranno belle le mie impronte digitali! Impronte anonime, senza identità, neutre come la plastica.
*[**** *ID_53 *fase_3 *genere_donna *areageo_africa *scrittura_singola *anno_2005 *autore_Scego]*

4. Discussioni

Negli studi sulle rappresentazioni sociali della migrazione nei media configurano la migrazione e i migranti sia come una minaccia alla sicurezza sociale sia attraverso un processo di vittimizzazione del migrante (Triandafyllidou, 1999; Mazzara *et al*., 2020; Bottura & Mancini, 2016). Allo stesso tempo, l'identità culturale del migrante è rappresentata come quella di un nemico, un invasore, un alieno, o un criminale (Binotto, 2015). I testi letterari analizzati propongono diverse costruzioni di senso riguardo la migrazione e l'identità culturale del migrante. Per quanto concerne quest'ultima, vi sono molti elementi di complessità. Affiora infatti una visione multipla dell'identità. In linea con la letteratura è presente una condizione di vittimizzazione legata allo sfruttamento lavorativo e/o sessuale, ma anche di resistenza rispetto al misconoscimento della propria identità. Inoltre, emerge una dimensione di normalizzazione della

quotidianità legata ad aspetti della vita di tutti i giorni, con narrazioni sulla natura, sui colori, sul cibo e/o sulle proprie emozioni. Infine, la scrittura sembra avere diverse funzioni, non solo di testimonianza e di resistenza, ma anche di espressività emotiva su aspetti legati alla vita quotidiana e di comunicazione del proprio punto di vista artistico-letterario.

5. Conclusioni

Queste analisi valorizzano il punto di vista delle scrittrici e degli scrittori africani e/o nati in Italia da genitori migranti di origine africana e mettono in luce rappresentazioni del migrante come clandestino, come bracciante, come vittima di sfruttamento e come individuo a cui è misconosciuta la propria identità di italiano nero. Affiora anche una dimensione della vita quotidiana, con narrazioni sulla cucina e sull'espressione di emozioni. Emerge quindi la figura del migrante come scrittore/scrittrice che attraverso la scrittura vuole lasciare un segno, di testimonianza e/o di resistenza da una parte, per mostrare una "normalizzazione" dell'esperienza quotidiana dall'altra. La ricerca psicosociale sui testi letterari, e in questo caso sulla letteratura italiana postcoloniale, può avere un importante valore pragmatico, al fine di prendere in considerazione il punto di vista di migranti o di italiani figli di migranti. Non soltanto la letteratura consente di dare spazio a interpretazioni in prima persona, ai protagonisti del percorso migratorio. Riteniamo le voci di scrittrici e scrittori siano particolarmente autorevoli, e generative, nel processo di costruzione e ricostruzione di conoscenze sociali condivise sulla migrazione e che metodologie di analisi quali-quantitativa quale quella qui presentata siano alleate preziose nei percorsi di ricerca.

Riferimenti bibliografici

Binotto M. (2015). Invaders, Aliens and Criminals: Metaphors and Spaces in the Media Definition of Migration and Security Policies. In Bond, E., Bonsaver, G., & Faloppa, F. (Eds.). (2015). *Destination Italy: Representing migration in contemporary media and narrative.* Peter Lang.

Bolasco S. (1999). *Analisi multidimensionale dei dati [multidimensional analysis of data]*. Roma: Carocci.

Bottura B. and Mancini T. (2016). Psychosocial dynamics affecting health and social care of forced migrants: a narrative review. *International Journal of Migration, Health and Social Care*, 12(2),109-119. https://doi.org/10.1108/IJMHSC-04-2014-0012

Contarello A. (2008). Social psychology and literature: Toward possible correspondence. In T. Sugiman, K. J. Gergen, W. Wagner, & Y. Yamada (Eds.), *Meaning in action: Constructions, narratives, and representations* (pp. 303–325). N.Y.: Springer. https://doi.org/10.1007/978-4-431-74680-5_18

Contarello A., Marini I., Nencini A. and Ricci G. (2011). Rappresentazioni sociali dell'invecchiamento tra psicologia sociale e letteratura. *Psicologia & Sociedade*, 23 (1), 171-180. https://doi.org/10.1590/S0102-71822011000100019

de Rosa A.S., Bocci E., Bonito M. and Salvati M. (2021) Twitter as social media arena for polarised social representations about the (im)migration: The controversial discourse in the Italian and international political frame. *Migration Studies*, 9(3), 1167-1194. https://doi.org/10.1093/migration/mnab001

Doise W. (1998). Social representations in personal identity. In Worchel, S., Morales, J.M., Páez, D., & Deschamps, J.C. (Eds.). (1998) *Social identity: International perspectives* (pp. 13-23). SAGE Publications Ltd,
https://dx.doi.org/10.4135/9781446279205.n2

Jacomella G. (2015). The Silence of Migrants: The Underrepresentation of Migrant Voices in the Italian Mainstream Media. In Bond, E., Bonsaver, G., & Faloppa, F. (Eds.). (2015). *Destination Italy: Representing migration in contemporary media and narrative*. Peter Lang.

Jodelet D. (1989) (Eds). *Les representations sociales.* Paris: Presses Universitaires de France.

Mancini T., Ceresini D. and Davolo A. (2007). Rappresentazioni sociali della società multiculturale. *Psicologia Sociale*, 1, 123-148. DOI: 10.1482/24339

Maneri M. (2011). Media discourse on immigration. The translation of control practices into the language we live by. In S. Palidda (Ed.), *Racial criminalization of migrants in the 21st century* (pp. 77–93). Ashgate.

Mannarini T., Veltri G. and Salvatore S. (Eds), (2020). M*edia and Social Representations of Otherness. Psycho-Social-Cultural Implications*. Cham, Switzerland: Springer Nature Switzerland.

Mazzara B.M., Avdi E., Kadianaki I., Koutri I., Lancia F., Mannarini T., Mylona A., Pop A., Rochira A., Redd R.E., Sammut G., Suerdem A., Veltri G.A., Verbena S. and Salvatore S. (2020). The Representation of Immigration. A Retrospective Newspaper Analysis. Journal of Immigrant & Refugee Studies, 19(4), 436-455. https://doi.org/10.1080/15562948.2020.1836299

Montali L., Riva P. and Frigerio A. (2013). The representation of migrants in the Italian press: A study on the Corriere della Sera (1992–2009). *Journal of Language and Politics*, 12(2), 226-250. https://doi.org/10.1075/jlp.12.2.04mon

Moretti F. (2013). *Distant Reading. London*; New York: Verso.

Moscovici S. (1976). *La psychanalyse: Son image et son public*. Paris: PUF.

Moscovici S. (1986). The Dreyfus affair, Proust and social psychology. *Social Research*, 53, 23-56.

Ratinaud P. and Marchand P. (2012). Application de la méthode ALCESTE aux « gros » corpus et stabilité des « mondes lexicaux »: analyse du « CableGate » avec IRAMUTEQ. 11e Journées internationales d'analyse statistique des données textuelles, Liège, Belgique (p. 835 844).

Reinert M. (1990). Alceste. Une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 26(1), 24 54.
https://doi.org/10.1177/075910639002600103

Romeo C. (2018). *Riscrivere la nazione. La letteratura italiana postcoloniale*. Mondadori, Milano.

Sbalchiero S. (2018). Topic Detection. A Statistical Model and a Quali-quantitative

Method. In A. Tuzzi (Ed.), *Tracing the Life-Cycle of Ideas in the Humanities and Social Sciences* (pp. 189-210). New York: Springer.
Triandafyllidou A. (1999). Nation and Immigration: A study of the Italian press discourse. *Social Identities: Journal for the Study of Race, Nation and Culture*, 5(1), 65–88. https://doi.org/10.1080/13504639951626
Triandafyllidou A. (2013). Migrants and the media in the twenty-first century. *Journalism Practice*, 7(3), 240–247. https://doi.org/10.1080/17512786.2012.740213
Tuzzi A. (2003). *L'analisi del contenuto: introduzione ai metodi e alle tecniche di ricerca*. Roma: Carrocci.

Alla ricerca del benessere: le famiglie adottive

Barbara Cordella, Alessandro Gennaro

University of Rome – barbara.cordella@uniroma1.it
a.gennaro@uniroma1.it

Abstract
In the article, we present research to explain how we use text mining in our field of study. The research, given as an example, was conducted with a group of adoptive mothers (37) interviewed on "the search for biological origins by their adoptive children". We carried out a cluster analysis and a subsequent correspondence analysis: 7 clusters and 6 factors emerged. Here we will consider only the top three factors, which represent more than 59% of the variance. Overall, the research shows the difficulty encountered by adoptive mothers in accompanying their children towards the rediscovery of biological origins. A subsequent regression model showed an associative link between the time elapsed since the adopted child's arrival in the family and the radicalism/reticence to change about the ways in which the issue is addressed, calculated in terms of the Euclidean distance of the subjects on the multifactorial space.

Keywords: Text mining, adoptive family, parental identity, recognition of diversity.

Riassunto
Nell'articolo presentiamo una ricerca per spiegare come utilizziamo il text mining nel nostro campo di studio. La ricerca, portata come esempio, è stata condotta con un gruppo di madri adottive (37) intervistate su"la ricerca delle origini biologiche da parte dei loro figli adottivi". Abbiamo realizzato una analisi dei cluster ed una successiva analisi delle corrispondenze: sono emersi 7 cluster e 6 fattori. In questa sede, considereremo solo i primi tre fattori, che rappresentano più del 59% della varianza. Nel complesso, la ricerca mostra la difficoltà incontrata dalle madri adottive nell'accompagnare i figli verso la riscoperta delle origini biologiche. Un successivo modello di regressione ha evidenziato un nesso associativo tra il tempo trascorso dall'arrivo in famiglia del bambino adottato e la radicalità/reticenza al cambiamento circa le modalità con cui viene affrontata la questione, calcolata in termini di distanza euclidea dei soggetti sullo spazio multifattoriale.

Parole chiave: text mining, famiglie adottive, identità genitoriale, riconoscimento delle diversità.

1. Introduzione

La funzione professionale di uno Psicologo della salute è quella di promuovere il benessere degli individui, dei gruppi, delle organizzazioni entro i singoli

contesti di vita. La traduzione di detta finalità (benessere) in obiettivi perseguibili, si avvale, spesso, di una rivisitazione (Cordella, 2016) della ricerca-intervento introdotta da Lewin (1946). In questi casi, in ragione di una determinata problematica, la ricerca ha carattere esplorativo e si propone di comprendere come, un certo gruppo di attori, si pone in relazione con uno specifico evento del contesto. È a partire da detta conoscenza, infatti, che è possibile progettare un possibile intervento.

La conduzione della ricerca, frequentemente, si avvale della raccolta di narrazioni (Freda, 2008), che vengono analizzate per rintracciare le dimensioni trasversali che le accomunano. Si assume, infatti, che il comportamento degli individui (anche verbale) sia orientato dalla cultura a cui si appartiene, continuamente prodotta nell'interazione ed espressa attraverso il linguaggio. In questa direzione, il linguaggio è considerato come lo strumento attraverso il quale si costruisce il legame sociale, ma è anche "il luogo" in cui si esprimono le dimensioni che conformano tale legame. Ne segue che, sul piano della ricerca, è possibile analizzare i testi prodotti da un determinato gruppo per coglierne le dimensioni culturali (trasversali alle singole narrazioni) che orientano la relazione, del gruppo preso in esame, con l'evento di interesse.

L'ampiezza del materiale raccolto e l'indubbia economicità in termini di tempo, rispetto alle procedure carta e matita, ci guidano all'utilizzo delle tecniche statistiche di text mining (Cordella, *et al*., 2014). Per aggregare la molteplicità dei dati qualitativi raccolti (forme grafiche) ci avvaliamo, dunque, di una analisi multivariata ed in particolare, di una un'analisi dei cluster, con algoritmo bisecting K-Means, limitato a 10 partizioni con l'esclusione delle unità di contesto che non presentano almeno una co-occorrenza, sulla matrice termini per unità di contesto. La partizione ottimale è individuata in relazione alla differenza del rapporto tra varianza between e totale della partizione n con quella (n-1). Successivamente sulla matrice termini per cluster, effettuiamo un'analisi delle corrispondenze.

A partire da dati qualitativi (forme grafiche), il nostro obiettivo è quello di definire un campo fattoriale, in cui collocare i cluster. Cluster e fattori assumono senso grazie alla co-occorrenza delle forme grafiche che consideriamo quale segno della cultura che si sta analizzando.

2. Un esempio di ricerca

Allo scopo di esemplificare quanto sopra indicato, presentiamo la ricerca condotta intervistando un gruppo di madri adottive invitate a narrare la propria prospettiva in relazione alla ricerca delle origini biologiche da parte dei loro figli adottivi. L'interesse per il tema nasce dalla constatazione che la letteratura si occupa del tema soprattutto considerando la prospettiva dei ragazzi adottati. Per l'esplorazione del tema è stata utilizzata una sola domanda aperta. Si è scelto,

inoltre, di introdurre una scheda per la raccolta dei dati anagrafici. La ricerca ha ottenuto l'approvazione del Comitato Etico del Dipartimento di Psicologia Dinamica, Clinica e Salute (protocollo n. 0000079 del 20/01/2020).

2.1. Partecipanti

Sono state condotte 37 interviste a madri adottive italiane, con una età compresa tra 46 e 64 anni (M=55,3; ds=4,7). Dalle schede per la raccolta dei dati anagrafici si osserva che il nostro gruppo di interviste comprende 15 adozioni nazionali e 46 adozioni internazionali; al momento dell'intervista da noi condotta, i figli avevano una età compresa tra i 14 e i 34 anni.

2.2. Analisi dei dati

Le interviste, integralmente registrate e trascritte, sono state lette per disambiguare le forme grafiche relative ai familiari (madre, padre, famiglia, nonni, ecc.), distinguendo i casi in cui ci si riferisce ai parenti biologici (segnati come bio) da quelli riferiti alla famiglia adottiva (segnati come ad). Inoltre, si sono sostituiti i nomi dei diversi luoghi di nascita con l'espressione "Paese di nascita". Successivamente, i testi sono stati raccolti in un unico corpus caratterizzato dai seguenti indici: numero complessivo di forme grafiche pari a 149501 (Tokens), di cui 10572 forme grafiche tra loro diverse (Type) e 5467 forme grafiche che si presentano una sola volta (Hapax). Il rapporto Type/Tokens è 0,071 e quindi inferiore a 0,1; il rapporto Hapax/Types è 0,517, ovvero prossimo al 50% e l'indice di Giraud che stima la complessità lessicale è 14,01. Infine, come osservato in figure1 i Type seguono una distribuzione Zipfiana.

Figura 1. Type distribution. Ranks in x axis and frequency in y axis

Constatata l'analizzabilità del testo e la sua ricchezza lessicale (Bolasco, 1999, Greco, 2016), la sua elaborazione, condotta con l'ausilio dello strumento T-Lab (Lancia, 2004), ha portato all'individuazione di 2789 contesti elementari. Dall'analisi risulta che la partizione ottimale si ha con7 cluster e 6 fattori. I primi tre fattori estratti, che spiegano il 59% del totale della varianza, sono stati interpretati come dimensioni di significato che organizzano il modo in cui i

genitori raccontano le richieste dei figli di avere conoscenza circa le loro famiglie di origine. Per ciascun partecipante è stato identificato il posizionamento sul piano multifattoriale definito dai tre fattori, ed è stata calcolata la distanza dall'origine del piano fattoriale in termini di distanza euclidea. Tale indice è stato interpretato in termini di polarizzazione individuale sulle dimensioni di significato descritte dai fattori: maggiore la distanza euclidea maggiore il peso di uno o più dimensioni fattoriali nell'organizzare il modo in cui viene attribuito senso alle richieste dei figli adottivi circa le proprie origini da parte dei genitori (per un approccio simile Gennaro *et al*., 2020; Gennaro *et al*., 2021). Implicitamente maggiore è la polarizzazione maggiore è la difficoltà a considerare altri modelli interpretativi delle richieste di conoscenza sulle proprie origini da parte dei figli. Infine un'analisi della regressione ha indagato il legame associativo tra il tempo trascorso dall'arrivo del bambino adottato in famiglia e la polarizzazione individuale sul piano fattoriale ricavato dall'analisi testuale. In tal modo è stato possibile stimare il ruolo del tempo nel definire una visione radicata/reticente verso possibili nuovi modelli di interpretazione della richiesta del figlio di conoscere la propria famiglia di origine. In secondo luogo una analisi della varianza ha indagato la differente distribuzione della polarizzazione nelle famiglie con una o più esperienze di adozione. Attraverso tale modello di analisi, infatti, è possibile stimare il ruolo della multi-adozione nel modulare le modalità con cui le richieste di conoscenza della famiglia di origine da parte dei figli vengono interpretate dai genitori adottivi.

2.3. Risultati e discussioni

Per ragioni di spazio, riduciamo la presentazione dei risultati ad una tabella (Tab. 1), che mostra le forme grafiche co-occorrenti nei primi tre fattori.

Considerando la co-occorrenza delle forme grafiche, è possibile individuare un primo fattore, da noi definito "La Famiglia", distinguendo tra il polo negativo che sembra connotare la famiglia come luogo in cui è avvenuto l'abbandono (la famiglia biologica) e il polo positivo che identifica nella famiglia un sistema che crea nuovi legami, che deve necessariamente riuscire a riparare la colpa dell'abbandono (la famiglia adottante), ma anche il rischio di non essere famiglia (infertilità). Il secondo fattore, invece, definito "L'estraneità", indica l'incontro con lo sconosciuto che, per definizione è non noto, fonte di pericolo, ed abbiamo distinto il polo negativo, come l'estraneità che spaventa a causa dell'oscillazione della propria identità, e il polo positivo come l'estraneità con cui ci si confronta a causa dei legami biologici. Il terzo fattore, segnato come le "Le difficoltà di integrazione", contrappone il polo negativo che narra la difficoltà scolastiche e, più in generale con il mondo esterno, alle difficoltà incontrate nella crescita del figlio adottivo, che richiede l'ausilio di esperti (psicologo). Nel complesso, lo spazio culturale che emerge dalle interviste disegna una identità genitoriale

precaria, continuamente impegnata nel riparare la propria storia, nel tentativo, mai esaurito, di ricondurla nei termini di una supposta normalità.

Tabella 1 Le forme grafiche co-occorrenti che caratterizzano i fattori.

Fattore 1				Fattore 2				Fattore 3			
La famiglia				*L'estraneità*				*Le difficoltà di integrazione*			
Che abbandona		*Deve riuscire*		*Oscillazione-paurose*		*Legami pericolosi*		*Mondo interno*		*Mondo esterno*	
Polo negativo		Polo positivo		Polo negativo		Polo positivo		Polo negativo		Polo positivo	
lemma	cont	lemma	cont	lemma	cont	lemma	cont	lemma	cont	lemma	cont
figlio bio	3,72	cominciare	2,92	paese di nascita	7,93	mamma bio	13,82	scuola	18,22	problema	12,58
abbandonare	2,57	riuscire	2,15	paura	4,47	fratello bio	9,55	classe	3,09	guardare	8,84
istituto	2,30	nonno ad.	1,14	tornare	4,20	parente bio	1,62	portare	2,09	aiutare	2,71
genitore bio	1,83	fatica	1,09	partire	2,83	tranquillo	0,36	lasciare	1,90	cercare	2,48
padre bio	1,70	rispondere	0,92	Italia	1,84	ospedale	0,30	capito	1,07	Figlio ad	2,25
famiglia bio	1,08	insieme	0,74	paese	1,53	internet	0,25	uscire	1,01	psicologo	2,09
rimanere	0,97	iniziare	0,39	ritornare	1,18	vero	0,23	Fratello ad	0,87	percorso	1,88
nonno bio	0,60	scrivere	0,31	passare	1,17	spiegare	0,19	roba	0,67	affrontare	0,66
realtà	0,55	mangiare	0,31	nascere	0,74	perdere	0,17	scuola	18,22	creare	0,62

Il modello di regressione condotto ha evidenziato risultati significativi (Tab. 2 e 3). Nello specifico l'esperienza adottiva, ovvero il tempo trascorso dall'adozione risulta essere un predittore significativo della polarizzazione individuale sul piano multifattoriale ricavato dall'analisi del testo (R=0,412; adj.R square 0,146; errore standard di stima = 0,066; $p<0,05$). Tale risultato indica il ruolo della dimensione temporale, quale fattore che tende a polarizzare il modo in cui i genitori attribuiscono senso alla richiesta di informazioni sulla propria famiglia di origine da parte dei figli adottivi. Ciò significa che man mano che i figli adottivi diventano adulti vi è una maggiore difficoltà ad interpretare le loro richieste inerenti le loro origini attraverso prospettive differenti.

Tabella 2. Anova test relativo al modello di regressione.

	Somma dei quadrati	**GdL**	**Media quadratica**	**F**	**Sig**
Regressione	0,031	1	0,031		
Residuo	0,154	35	0,004	7,155	0,011
Totale	0,186	36			

Tabella 3. Modello di regressione con polarizzazione come variabile dipendente ed il tempo di esperienza adottiva come variabile indipendente.

	B	**Errore standard**	**Beta**	**T**	**Sig**	**VIF**
(Costante)	0,088	0,029		3,005	0,005	
esperienza adottiva	0,000	0,000	0,412	2,675	0,011	1,000

Infine l'analisi della varianza (Tab. 4) ha evidenziato come i genitori con una sola esperienza di adozione hanno livello di radicalità/reticenza inferiore rispetto ai genitori con più esperienze di adozione.
Alla luce di tale dato potremmo sostenere che l'esperienza multi-adottiva espone i genitori ancor più all'estraneità dei figli e tale esposizione si traduce in una maggiore radicazione nei propri canoni culturali di interpretazione delle loro richieste.

Tabella 4. Confronto polarizzazione tra famiglie con esperienza adottiva singola e multipla.

esperienza di adozione	N	Media polarizzazione	dev.st.	errore standard	somma dei quadrati	F	sig
singola	16	0,1331	0,0616	0,0154	0,021	4,384	0,044
multipla	21	0,1807	0,07346	0,01603	0,165		

2.4. Conclusioni

L'analisi dei risultati raggiunti suggerisce almeno due riflessioni. In una prospettiva più metodologica, il lavoro evidenzia la possibilità di utilizzare strumenti e metodologie per l'analisi testuale come mezzo per indagare i framework culturali e simbolici uscendo dalla dicotomia qualitativo-quantitativo che molto spesso ha limitato lo studio e l'analisi dei trascritti in diversi campi del sapere. In una prospettiva più clinica, inerente al lavoro di ricerca riportato, i risultati evidenziano la necessità per i servizi clinici di spostare il focus dei loro interventi da una visione di promozione dell'accoglienza del figlio adottato ad una visione di promozione dell'accompagnamento delle famiglie adottive nel riconoscimento dell'estraneità dei propri figli come volano per l'integrazione dei processi di sviluppo familiare.

Bibliografia

Bolasco S. (1999). *Analisi multidimensionale dei dati: metodi, strategie e criteri di interpretazione* [Multidimensional data analysis: methods, strategies and criteria of interpretation]. Roma: Carocci.
Cordella B. (2016). La ricerca intervento in psicologia clinica. In Grasso M., Cordella B., Pennella A.R., *L'intervento in Psicologia Clinica*. Carocci.
Cordella B., Greco F., 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, S. Fleury (Eds), *Proceedings JADT 2014, 12es Journées internationales d'Analyse Statistique des Données Textuelles*, Paris, France, Juin 3-6, pp.173-184.
Freda M.F. (2008). *Narrazione e intervento in psicologia clinica. Costruire, pensare e trasformare narrazioni fra «logos» e «pathos»*. Liguori.
Gennaro A., Carola V., Ottaviani C., Pesca C., Palmieri A., Salvatore S. (2021). Affective Saturation Index: A Lexical Measure of Affect. *Entropy* 2021,23,1421.

https://doi.org/10.3390/e23111421
Gennaro A., Kypp S., Viol K., De Felice G., Andreassi S., Salvatore S., Schiepeck G. (2020). Phase Transition of the Unconscious: Text analysis of dream in psychoanalytic psychotherapy. *Frontiers in psychology* (11):1667.doi.org/10.3389/fpsyg.2020.01667
Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale.* Franco Angeli.
Lancia F. (2004). *Strumenti per l'analisi dei testi. Introduzione all'uso di T-Lab*. Franco Angeli.
Lewin K. (1946). Action Research and Minority Problems. *Journal of Social Issues* (2): 34-46.

The 'words' of No green pass communities on Twitter. A two-mode semantic network analysis

Rosario D'Agata[1], Domenico De Stefano[2], Francesco Santelli[3]

[1]University of Catania – rodagata@unict.it
[2]University of Trieste – ddestefano@units.it
[3]University of Napoli Federico II – fsantelli@unina.it

Abstract
The most extreme positions against anti pandemic government measures are concentrated in the most democratic countries. Therefore, on the one hand, freedom of speech is guaranteed, being it the essence of modern democracy; on the other stands the awareness of the risks that this freedom implies for the entire community. We are facing what some scholars call "the dilemma" of democracies. Increasing the drastic consequences of this dilemma is undoubtedly the communicative power of social networks. In this work, starting from data collected on Twitter from 25-Sep-2021 to 22-Oct-2021 (that is the period when local elections in many important cities of Italy took place) concerning the green pass debate in Italy, we construct a two-mode Semantic Network, which is a bipartite graph that describes connections between two types of nodes, social actors (in our case twitter users) and semantic concepts. In particular, in the first step data-tweets size will be significantly reduced; in the second step, a content analysis will be run to pinpoint the concepts underneath them. To detect communities of users and concepts, we use a proper two-mode community detection approach, i.e. an extension of the fast-greedy suited for the bipartite network called "DIRTLPAwb+". The aim is to identify communities of users expressing different opinions and concepts within the green pass debate.

Keywords: two-mode semantic networks, two-mode community detection, Twitter, green pass

1. Introduction

Historically, positions against vaccinations have always existed. The propagation of anti-vaccine messages, however, took place through limited means of information. In the 1800s the debate involved the educated social classes. Most of the population was illiterate and was reached by anti-vaccine messages through the use of cartoons that did not require particular efforts to deepen. From a more strictly political-institutional point of view, the contexts in which such

debates developed were far from modern democracies. Widespread illiteracy and the lack of democratic institutions contributed to the success of vaccination campaigns. Paradoxically, as Wilson and Wiysonge (2020) have pointed out, the most extreme vaccine positions are concentrated precisely in the most democratic countries with higher levels of education. On the one hand, therefore, the guarantee of freedom of speech, which is the essence of modern democracy; on the other, the risk that this freedom could lead to dangerous behavior for the entire community. We are faced with what the authors call "the dilemma" of democracies. At increasing the drastic consequences of this dilemma is undoubtedly the communicative power of social networks.

The role of social media in exacerbating vaccination hesitancy has long attracted the attention of scholars not only in the sociological field but also in the medical and epidemiological fields (Davies *et al*., 2002; Keelan *et al*., 2007; Evrony and Caplan, 2017). Unlike other manifestations related to the so-called 'post truth', of which no vax speeches are an expression (Kata, 2010), the effects on the health of communities due to the propagation of messages against vaccinations have a more serious impact. Many studies relate the spread of 'no vax' messages with behaviors of rejection of vaccination practices with the consequent spread of epidemics (Herrera-Peco *et al*., 2021; Ball, 2020; Chou *et al*., 2018; Larson, 2018). This last aspect does not escape the WHO which sees precisely the spread of no vax groups on social networks as a threat to global health in the coming years (WHO, 2019).

About the content of 'no vax' messages on social networks, it is possible to identify a scale that from the general, clear opposition to any form of vaccination, reaches the particular, opposition to the single measure aimed at containing the pandemic (e.g. lockdown, masks, green pass, etc.). In an interesting work, Al-Ramahi *et al*. (2021) analyze the content of about 51,000 posts in English published on Twitter between January and October 2020 and related to the use of masks by identifying 10 main themes that the authors summarize in 3 macro areas. The first, which we can define as political, contains those semantic elements connected to the concept of 'constitutional rights' and 'freedom of choice'; the second, of a conspiratorial type, recalls the conspiratorial idea of control over the masses by 'powerful' subjects as well as the role of Big Pharma and finally the third, of pseudo-scientific matrix, is based on the publication of news with formally scientific content but based on incorrect data or directly on fake news and aimed at asserting the non-existence of the pandemic.

In this work, we adopt a two-steps procedure to extract the semantic concepts from the tweets and build up the two-mode network to analyze. In particular, the first step consists in reducing the data dimensionality of the tweets, and in the second step, we perform a content analysis to pinpoint the concepts they express. To detect the communities of users that are linked via the connections they have with the

concepts we use a proper two-mode network community detection approach. The aim is to identify communities of user expressing different opinions and concepts (in terms of hashtags) within the green pass debate in the period of the 2021 Italian administrative elections which exacerbated the debate on this topic.

2. Methodology and network framework

The network that we define is a two-mode (or bipartite) network. It means that nodes, primary elements of a network, are of two distinct types and that connections among elements of the same type are not allowed. Thus, links exist only connecting nodes of one type to nodes of the other type. In this case, *hashtags* and *users* are the two distinct sets of nodes.

Links among nodes are defined as not directed, with symmetric relationships. Further, such links are conceived as weighted, with weights equal to the number of times the given user uses the hashtag. In this framework, we are interested in the patterns of such weighted links, to discover consistent groups of both hashtags and users, in order to provide the interpretation of the semantic content and of the relationships between the two sets of nodes. For this purpose, the core of the analysis will be on the *Community Detection* phase.

Community Detection (see e.g. Newman, 2013), aims to identify the presence of groups of nodes that are tightly connected in knit groups within communities and loosely connected with hashtags and users belonging to other communities. So, nodes are densely connected within the community, and less with members of other communities. Community detection in one mode network is based on Newman's modularity (Newman, 2013). We propose, instead, the application of a two-mode community detection (or, equivalently here, a community detection for bipartite graph). Such extension is implemented in the algorithm DIRTLPAwb+ (Beckett, 2016) proposed originally by Barber (2007). The algorithm initializes all nodes to be in their community, for a total of communities equal to the number of nodes. The modularity is conceived in bipartite graph as follows:

$$Q_b = \frac{1}{m}\sum_{u=1}^{r}\sum_{v=1}^{c}\left(\tilde{A}_{uv} - P_{uv}\right)\delta(g_u - h_v) = \frac{1}{m}\sum_{u=1}^{r}\sum_{v=1}^{c}\left(\tilde{A}_{uv} - \frac{k_u d_v}{m}\right)\delta(g_u - h_v) \quad (1)$$

In this formulation, P is the null model matrix representing the probability of links between nodes, given the degree distributions of $\tilde{A}$, the block diagonal affiliation matrix with r nodes of the first type and c nodes of the second type; k, is the node degree for 1st type nodes while d is the node degree for 2nd type nodes.

1st type node labels are denoted by g, while h are the labels for 2nd type nodes, and the Kronecker delta function $\delta(g_u,h_v)$ is equal to one when nodes u and v are classified as being in the same community or zero otherwise. In this contex, we denote as affiliation matrix che representation of the graph: it has as much rows as number of nodes of the first type and as much columns as nodes of the second type.

3. Data description

Data have been collected via Twitter API and its interface with R (R version 4.1.3) and Rstudio (version 1.4.1103). Such free API, used mainly by means of the packages "rtweet" and "twitteR", are able to search and download tweets and their related meta-data linked to a specific hashtag. In this case, our query has been "#nogreenpass", so all the included status have to include such hashtag. Futher, the specification of the language has been Italian, since we are interested to deepen the Italian debate about the topic. The window range of the collected tweets goes from 25-Sep-2021 to 22-Oct-2021. In this month of data collection, the total number of tweets is 31641. With the usage of the free tools available, is it not possible to obtain a random sample of the tweets but instead a selection of it that is not necessarily random. By the way, enlarging the time window it is possible to obtain a good proxy of the overall debate on such social platform.
The dataset contains only original status, i.e. when a user decide to post a new status (a new tweet) on its profile. Retweets and replies have instead been excluded. Retweets, for their own nature, give no additional information, in the semantic field, with respect to the original status. It can be used of course to analyze dimensions such as engagement or virality of concepts/tweets/users, but such analysis are not in the scope of this contribution. Replies could be very useful in analyzing how opinion spread in the online platforms as a signed network structure (see e.g. De Stefano and Santelli, 2019), but they assume a different approach in their analysis because most of the time the concept they express is conditional to the original sentiment/topic shown in the status. We seek, instead, for a one-level graph in which all the nodes are on the same level, such in the original statuses. Users names are left untouched, while hashtags are all put to lower case to have more consistent results.

Table 1. Brief description of the larger communities (≥10 unique users and hashtags) among the 28 detected. The community label is dervied from the hashtags with the largest number of occurrences. Number of users and hashtags is the number of unique of users and hashtags within each community

Community label	**N° of users**	**N° of hashtags**
1 Generic protests against the green pass	21	13
2 Solidarity with port workers	8	7
4 Popular Tv talk shows	3	18
5 International protests (France)	23	26
7 Trieste	40	18
11 Against the Government	9	22
12 No mandatory vaccine	6	15
13 Cliché about politics	1	12

15 Right to work.	5	19
17 Against extreme Right-wing party	4	9
19 TV show (Rete 4)	3	15
20 General Strike	9	25
23 No Vax	15	9
25 Against Draghi	12	15

In our data, the total number of used hashtags is 7880, while the total number of users is 6671. The resulting graph has, thus, 14551 nodes and 41253 links, undirected and weighted, among them. In the spirit of bipartite graphs, connections are build only between nodes of different types, i.e. links connect users and hashtags. From further analysis, the hashtag "#nogreenpass" is excluded, given that it is the query of the data collection phase.
Such graph is a large and sparse graph (sparsity of the adjacency matrix is over 99% and edge density is < of 0.0002). About 50% of hashtags are used just one time, and about 50% of users appear in the data only one time well. Overall, to obtain a certain amount of statistical information from the network, we decided to prune such graph deriving an induced subgraph that contains only the top 3.3% of nodes in terms of degree. The final graph includes 485 nodes (282 hashtags and 203 users), and the edge density has increased to 0.0063. In this way, we analyze a network in which hashtags has been used at least 10 times in the period considered, seeking for communities with relevants hashtags, that are a proxy of the topics, and relevant users, that posted several times using the hashtag "#nogreenpass".

4. Preliminary results and future developments

In the procedure, the algorithm identifes as best solution the case with 28 communities. The communities are qualitatively described according to the most frequent hashtags. In Table 1 we report only some the larger communities (thos with more than 10 unique users and hashtags) among the 28 detected. The network representation of the community structure is depicted in Figure 1. Most communities are polarized on different aspect of the protest against the green pass or other covid-19 containment measures. Several communities have a quite small size in terms of unique user and hashtag counts. Not all of the identified communities are composed of users aligned toward the protest (for instance community 17). The largest community which is quite interesting for the users involved is focused on the opponents of Italy's green pass prostest held in the city of Trieste (community 7, in which the most used hashtag is #Trieste with 3964 occurrences).

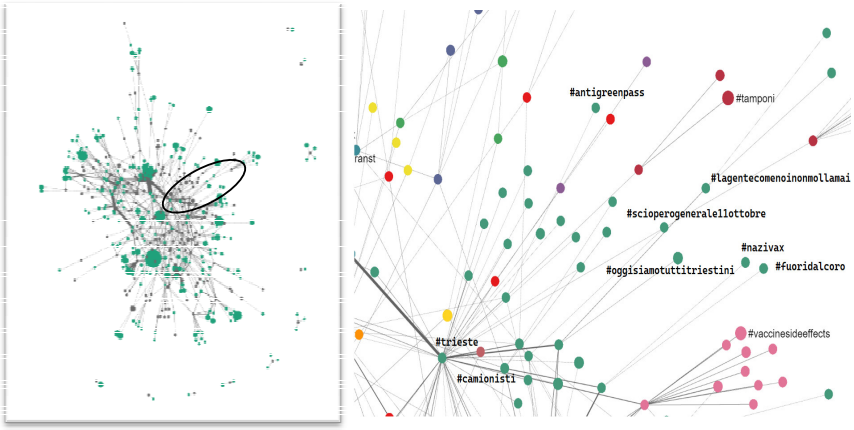


Figure 1. On the left panel, graphical representation of the two-mode network of users-hashtags in green hashtags and in gray users (in the black circle nodes of the community 7). On the right panel, zoom on the community 7 (members are depicted in green).

Community 7 involves the most extreme protesters against the green pass. This is confirmed by the distribution of the used hashtag (e.g., #fuoridalcoro with 192 occurrences; #lagentecomenoinonmollamai, 124 occurrences; #nazivax, 98; see Figure 1, right panel). Preliminary results are quite promising in identifying groups of users sharing similar (sometimes extremes) opinions. In the future developments of the proposed approach we will deepen the analysis of the identified communities by investigating their internal structures by means of appropriate individual network measures and eventually ego-network indices. In this way we will also be able to identify how the prominent nodes in the green pass debate behave in other current trending issues on the Twitter platform.

References

Al-Ramahi M., Elnoshokaty A., El-Gayar O., Nasralah T. and Wahbeh A. (2021) Public Discourse Against Masks in the COVID-19 Era: Infodemiology Study of Twitter Data. *JMIR Public Health Surveill*, 7(4).

Ball P. (2020) Anti-vaccine movement could undermine efforts to end coronavirus pandemic, researchers warn. *Nature*, 581 (251).

Barber MJ. (2007) Modularity and community detection in bipartite networks. *Phys. Rev. E* 76 (6).

Beckett SJ. (2016). Improved community detection in weighted bipartite networks. *R Soc Open Sci.*, 3.

Chou W.Y.S., Oh A. and Klein W.M.P. (2018) *Addressing Health-Related Misinformation on Social Media. JAMA* 2018, 320(23): 2417–2418.

Davies P., Chapman S. and Leask J. (2002). Antivaccination activists on the world wide web. *Arch Dis Child*. 87 (1): 22–25.

De Stefano D. and Santelli F (2019). Combining Sentiment Analysis and Social Network Analysis to Explore Twitter Opinion Spreading. *28th International Conference on Computer Communication and Networks* (*ICCCN*), 1-6.
Evrony A and Caplan A. (2017) The overlooked dangers of anti-vaccination groups' social media presence. *Hum Vaccin Immunother*, 13(6):1-2.
Herrera-Peco I., Jiménez-Gómez B., Romero Magdalena C.S., Deudero J.J., García-Puente M., Benítez De Gracia E. and Ruiz Núñez, C. (2021) Antivaccine Movement and COVID-19 Negationism: A Con-tent Analysis of Spanish-Written Messages on Twitter. *Vaccines (Basel, 9(6)*: 656.
Kata A. (2010). A Postmodern Pandora's Box: Anti-Vaccination Misinformation on Facebook. *Vaccine*, 28,1709-1716.
Keelan J., Pavri V., Balakrishnan R. and Wilson K. (2010). An analysis of the human papilloma virus vaccine debate on MySpace blogs. *Vaccine*, 28(6) :1535-40.
Lancichinetti A. and Fortunato S. (2009). Community detection algorithms: A comparative analysis. *Physical Review* E,80 (5).
Larson H.J. (2018) The biggest pandemic risk? Viral misinformation. *Nature*, 562(309).
Newman M. E. J. (2013). Community detection and graph partitioning. *Europhysics Letters*, 103 (2)
Wilson S.L. and Wiysonge C. (2020). Social media and vaccine hesitancy. *BMJ Global Health,* 5, 1-7.
World Health Organization (2019) Ten Threats to Global Health in 2019.

Sentiment analysis of emotions in social media during the COVID-19 pandemic

Angela Maria D'Uggento[1], Albino Biafora[2], Claudia Marin[1], Fabio Manca[1], Massimo Bilancia[1]

[1]University of Bari (DIEF) – angelamaria.duggento@uniba.it claudia.marin@uniba.it fabio.manca@uniba.it massimo.bilancia@uniba.it
[2]Polytechnic of Bari (Master's Degree in Data Science) – a.biafora50@studenti.poliba.it

Abstract
In recent decades, social media has attracted increasing interest from researchers as a valuable source of data for analyzing users' opinions on specific topics. When extreme events occur, netizens' comments spread quickly online and can influence people's judgments both positively and negatively. Online social media such as Twitter provide a platform for netizens to learn about epidemics and express their opinions. Therefore, social media platforms can be a good tool for governments, businesses, and other organizations to learn about and respond to public concerns and emotional reactions with institutional messages or emotional comfort. During the Covid 19 pandemic, many companies used television commercials to build trust and hope in Italian families by emphasizing the principles of solidarity and fraternity and putting the brand in the background. In this paper, we attempt to measure the effectiveness of a change in the advertising approach of big brands using a specially created corpus of 20,982 tweets. Our results show that Internet users expressed an overall sense of belonging and gratitude during the lockdown, as well as an appreciation for television advertising that was perceived as disruptive at other times. We also became aware that the only currently available Italian lexicon for text analysis, developed by CNR, needs to be improved because it deals with formal Italian vocabulary and is not suitable for informal language such as tweets. The decoding of irony in a text by automatic polarity determination techniques remains a problem that needs further study.

Keywords: Sentiment analysis; Opinion mining; COVID-19 pandemic.

1. Introduction

In 2020, the planet was struck by an unknown pandemic that had unexpected and devastating effects and is still ongoing, reappearing in a second, third and even fourth wave in different parts of the planet. This led to radical changes in

the way individuals and communities lived, organized their work, studied, and spent their leisure time. Due to a long lockdown, the world's population was forced to reorganize their routines, adjust their habits, and change their perception of what was previously considered "normal". The coronavirus emergency outlined a critical situation, not only from a health perspective, but also from an economic and social perspective. Unemployment, increasing inequalities and crises in numerous sectors such as tourism, are just some of the dramatic effects caused by the ongoing pandemic. Companies have adopted different strategies depending on the type of goods sold, but with common objectives, such as respecting the sensitivity of many people affected by the sudden loss of a loved one and increasing empathy with consumers. During the period of total closure of all production units, companies have taken two different paths: some of them have simply closed their physical premises and waited for the possibility of reopening, while the others have used the available resources to maintain operations, changing their objectives in the short term or simply changing the channels to reach them. This second group has turned to a large cluster of consumers who, deprived of everyday life, looked for it on the Internet: Restaurant dinners were replaced by food delivery or orders from home; shopping afternoons were replaced by frequent e-commerce purchases, and face-to-face relationships were replaced by video calls and social media interactions. In Italy, social media use increased by 21% after the start of the lockdown compared to the end of 2019 (StatCounter, 2022). This puts Italy in second place after Spain, where the increase was 48% (Pérez-Escoda *et al*., 2020).

Under these premises, this paper aims to analyze the impact of these new paradigms presented in some commercials on TV during the first Italian total lockdown, from February to May 2020, and the reactions they provoked among consumers. The analysis of thousands of tweets published during this period allowed to verify the effectiveness of the advertising message contained in the TV spots and to highlight the awareness of the message itself. For this research, sentiment analysis techniques (Aggarwal and Zhai, 2012) were used to analyze a large number of tweets collected over the Internet. This approach, discussed in more detail in Section 2, complements traditional methods and can be seen as a proposal to use the most advanced data analytic techniques to measure an advertisement's ability to capture customers' attention.

2. Methods

2.1. Sentiment analysis

Sentiment analysis (SA) aims to find opinion-forming information and detect the "feeling" in texts (Pang and Lee, 2008). It can be considered as an evolution of text mining, which is a knowledge-intensive process that deals with information retrieval, text analysis, pattern and terminology extraction, document

classification, clustering, visualization, data mining, and machine learning (Dang and Ahmad, 2015; Lee and Park, 2018). Specifically, SA is the computerized study of people's thoughts, opinions, feelings, evaluations, attitudes, impressions, moods, and emotions. In this context, emotions such as approval, admiration, dislike, shame, and contempt can be identified and used to create predictor variables. It is now one of the most active research areas in natural language processing (NLP) and its applications affect many fields, starting with economics and social sciences, due to its importance in business and society. In recent years, as more and more people use social media to comment on events of all kinds and express their opinions, an immense amount of textual data has become available for analysis by researchers. Any kind of text published on the Internet, whether structured or unstructured, can be considered as a corpus that can be analyzed using the methods of SA. Following Liu and Zhang (2012), we use a lexicon-based approach in this work, where a document is scored based on a sentiment lexicon by summarizing the sentiment scores of all words in the document. This approach contrasts with machine learning algorithms for SA, which typically rely on supervised classification methods, often in conjunction with deep neural transformer models to support language understanding and decoding of emotional content (Devlin at al., 2019). For comprehensive overviews of the methods involved with SA see Liu and Zhang (2012), Tsytsarau and Palpanas (2012), and Giachanou and Crestani (2016).

Statistical models dealing with SA focus on popularity (positive, negative, or neutral), as well as emotions (joy, anger, frustration, sadness, etc.) and/or expressions of interest (e.g., interested – not interested), and also track trends over time: did the launch of a new product or the day a particular press release occurred trigger positive or negative sentiment? Companies are then faced with the task of analyzing huge amounts of text data (emails, tickets, chats, conversations on social networks, surveys, articles, documents, etc.), which would be impossible without special techniques. Once the text data is embedded in a suitable space, it is even possible to perform real-time analysis: SA, in this case, aims to detect issues live on social media, and it is possible to immediately understand if a customer is angry, disappointed, happy, etc. Thus, it becomes possible to intervene and evaluate the sentiment towards the brand or product.

2.2. Data sources and preparation

Twitter was founded in 2006 and enjoys an ever-growing number of users thanks to the extremely easy way to retrieve, post and download user comments. It is considered one of the largest collections of data on user-generated content. Twitter Sentiment Analysis (TSA) is a special subsection of SA, created to analyze the sentiments expressed in messages posted on the platform. Given the informal nature of the communication and the length constraints, the language

of tweets is very different from the language used in other types of text (web, blogs, news). The most common peculiarities in tweets are emphasized capitalization, emphasized elongation, abbreviations, and the use of slang and neologisms, all of which have a strong impact on the tokenization step.
To analyze the entire corpus of collected tweets, we created two separate lexicons for positive and negative terms, respectively (Kiritchenko *et al*, 2014). The Python programming language was chosen for the extraction of the tweets, the creation of the dataset, and the subsequent preprocessing phase. Python has several libraries to extract tweets directly from the social network, but not all of them allow to extract older tweets. The library GetOldTweets3 was used to extract tweets published from 2020/01/01 to 2020/06/30, using as a search query a list of words related to the goals of the paper. The list of search terms was generated by searching for the brands appearing in the commercials, the hashtags they launched, and more general terms such as "TV commercial" or "coronavirus TV commercial" (for the full list of search terms see Table 1 below). The following fields were extracted: Date of tweet, ID of tweet, original tweet, search key for extraction, and reference tag. A total of 20,982 tweets in Italian were collected, stored in a .csv file with UTF-8 encoding and prepared for the preprocessing phase. In this step, tweets were tagged with labels referring to three pre-defined time intervals: before the lockdown, from 2020/01/01 to 2020/03/08; during the lockdown: 2020/03/09 to 2020/05/03 (the so-called "phase 1"); after the lockdown: from 2020/05/04 to 2020/06/30. The cleaned tweets, with the stop words removed, constituted the text corpus on which the study was conducted and were integrated with the metadata: ID, date and time, search key, reference tag and the marker referencing the time period.

Table 1. Search terms for the tweet: brand names, hashtags and keywords

```
'Barilla','#Barilla','#ItaliaCheResiste','MulinoBianco','Mulino Bianco', 'Coop',
'#Coop','#MulinoBianco','Lavazza','#Lavazza','Bauli','#Bauli','#PasquaConITuoi','Esselunga'
,'#Esselunga','Conad','#Conad','Carrefour','#Carrefour','Famila','#Famila','#InsiemeCeLaFar
emo','Axa','#Axa','BMW','#BMW',#InsiemePerProteggerci','#InsiemePerRipartire','Head&Shoulde
rs','FCA','#FCA','#Head&Shoulders','#NoiCiSiamo','Mazda','#Mazda','Toyota','#Toyota','#Ripa
rtiremoInsieme','Renault','#Renault','Volkswagen','#Volkswagen','#IoRestoInGarage','Pupa','
#Pupa','#TorneràIlSorriso','#CondividiUnSorriso','WindTre','#WindTre','Chiquita','#Chiquita
','CocaCola','Coca-Cola','CocaCola','#Coca-Cola','SpotTv',
#SpotCovid','SpotCoronaVirus','#SpotTv','SpotPubblicitario','#SpotPubblicitario','SpotPubbl
icitari', '#SpotPubblicitari','SpotCovid19','#SpotCovid19','SpotCovid','Spot Corona
Virus','Spot Lockdown', 'Pubblicita Lockdown','Spot Tv Covid','Spot Tv Corona
virus','Pubblicità CoronaVirus','Pubblicità Corona Virus'.
```

3. Results and discussion

3.1. Analysis of sentiment

Sentiment analysis involves quantifying sentiment in the extracted corpus and assigning a degree of positivity/negativity or neutrality toward the topic to each document (tweet). However, a caveat is necessary: While there are numerous

lexicons in English, such lexicons are very scarce in Italian. The only available Italian resource, the opeNER lexicon (Russo *et al*, 2016), was developed by the National Research Council (CNR) and kindly provided by the authors. The use of a lexicon is one of the main approaches to sentiment analysis and involves the calculation of sentiment based on the semantic alignment of words in a text. After tokenizing each tweet, the first step is to score each term by assigning a polarity (negative, positive, or neutral) and a score using the lexicon mapping. An initial general analysis showed that 58% of all tweets were positive, while 25% of tweets were classified as negative. Thus, consumer comments on Twitter about TV brand ads appear to be positive for more than half.

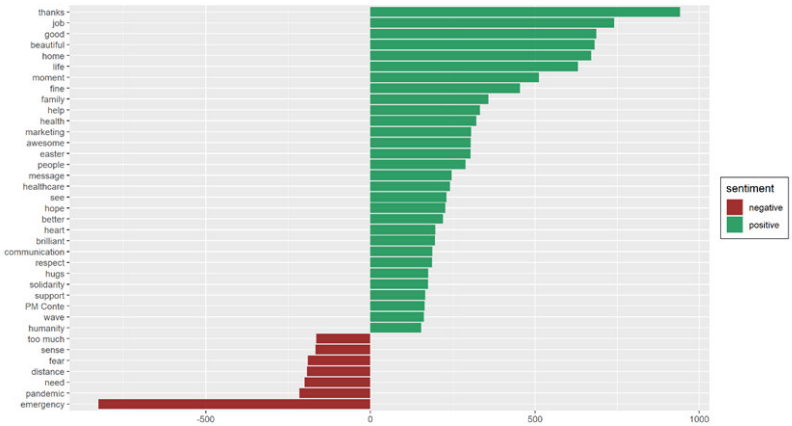


Figure 1. Barplot of the contributions to the general sentiment divided into negative and positive polarity.

An analysis stratified by the time period showed that the percentage of positive tweets during the lockdown was 61%, while it was lower in the pre- and post-lockdown periods, albeit by only 7%. To better understand the composition of the overall tweet dataset and make a comparison between terms, it is useful to look at the Figure 1 above, which shows the contribution of a given term to the calculation of overall sentiment. This result was obtained by multiplying the score of each term (positive or negative) by the number of occurrences of that term in the 20,982 tweets

To analyze sentiment and examine its trend during the three study periods considered, the daily average was calculated by adding the scores of the terms within the same tweet to obtain the total rating of the tweet, and then dividing all the ratings of each tweet posted on the same calendar day by the number of tweets for the day. The average of the overall sentiment was 0.41 with a standard deviation of 1.0, and the main statistics of the scores of each tweet in the three

periods are as follows:

Table 2. Summaries of the scores of the individual tweets.

Period	min	max	mean	sd
Before lockdown	-4.000	3.833	0.271	0.932
During lockdown	-4.750	7.750	0.501	1.013
After lockdown	-3.833	6.000	0.333	0.999

The trend in average daily sentiment scores is shown in Figure 2. It shows a dip in average scores just before the lockdown, when the highest negative average score was reached, a sharp recovery in positive sentiment during the lockdown phase, and a drop just after the end of phase 1, when people returned to a normal life.

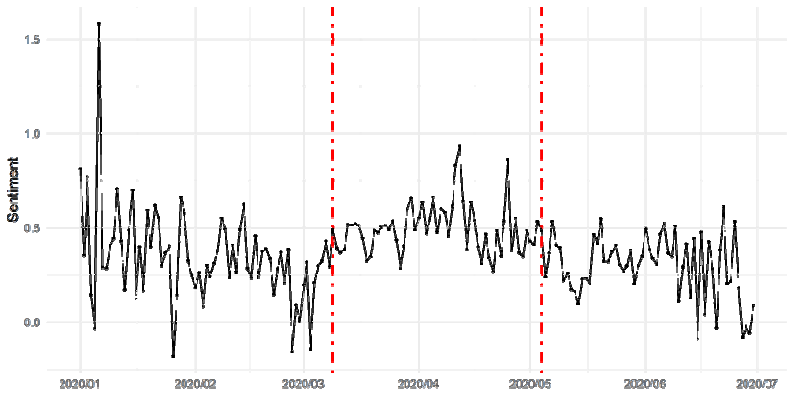


Figure 2. Trend of the average daily sentiment score.

5. Discussion and conclusions

The main results of this work confirm that the major brands changed their communication strategy and based their commercials TV on the principles of solidarity and fraternity, in order to increase trust and hope in Italian families during the pandemic response. The formats of their commercials were inevitably changed and triggered responses from customers, which can be measured by the increasing number of tweets and their content, as well as by the digital daily activities. Text analytics arises from the need to automate the understanding of unstructured textual data found not only in books and newspapers, but also in posters, blogs, customer reviews on the Internet, and posts on social media. Our proposal addresses an ad hoc corpus of unstructured data and transforms it to perform text mining analysis and identify patterns and trends. After a careful

phase of extracting and pre-processing tweets with natural language processing algorithms, a dataset of 20,982 tweets was created and analyzed, considering three different time periods: before the lockdown, during the lockdown, and after the lockdown. Thus, the novelty of the brand's advertising approach seems to have evoked, above all, a sense of gratitude and belonging that is not present in the other two periods, in which it was found that the spot was "tweeted" mainly for the brand's product or as a disruptive element.

In subsequent research, we aim to develop new positive contributions to overcome the limitation imposed by the lexicon and computational libraries (for both R and Python), which are rich for English but poor in content and often incomplete for Italian. The translation of Italian corpora into English is not trivial: a hard preliminary work consisted in performing the text analysis by constructing a suitable string similarity function, where the lexicon contains only regular word forms, and the conjugation of verbs is another non-negligible difference. The analysis of bi-grammes and co-occurrences could improve the construction of the Italian lexicon and, in view of the objectives of this study, help to identify the most effective spots and their features.

References

Aggarwal C.C. and Zhai C. (2012). *Mining Text Data*. Springer Nature. doi:10.1007/978-1-4614-3223-4.

Dang S. and Ahmad. (2015). A Review of Text Mining Techniques Associated with Various Application Areas. *Int J Sci Res* (IJSR), 4(2), 2461-2466.

Devlin J., Chang M.W., Lee K., and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, pp. 4171–4186. doi:10.18653/v1/N19-1423.

Giachanou A. and Crestani F. (2016). Like it or not: a survey of Twitter Sentiment analysis methods. *ACM Comput. Surv*, 49(2):Article 28. doi:10.1145/2938640.

Kiritchenko S., Zhu X. and Mohammad S. M. (2014). Sentiment analysis of short informal texts. *J. Artif. Intell. Res.*, 50, 723–762. doi:10.1613/jair.4272.

Lee Y. and Park J. (2018). Identification of future signal based on the quantitative and qualitative text mining: a case study on ethical issues in artificial intelligence. *Qual Quant*, 52(2), 653–667. doi:10.1007/s11135-017-0582-8.

Liu, B. and Zhang L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In Aggarwal C. and Zhai C. editors, *Mining Text Data*, Springer Nature, pp. 415-463.

Pang B. and Lee L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2), 1-135, doi:10.1561/15000000011.

Pérez-Escoda A., Jiménez-Narros C., Perlado-Lamo-de-Espinosa M., Pedrero-Esteban L.M. (2020) Social Networks' Engagement During the COVID-19 Pandemic in Spain: Health Media vs. Healthcare Professionals. Int *J Env Res Public Health*, 17(14):5261. doi:10.3390/ijerph1714526.

StatCounter GlobalStats (2022). https://gs.statcounter.com/social-media-stats/all/italy/2020, retrieved on 2022/03/14.

Russo I. *et al*. (2016). OpeNER Sentiment Lexicon Italian - LMF, ILC-CNR for

CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council (CNR), Pisa. http://hdl.handle.net/20.500.11752/ILC-73.
Tsytsarau M. and Palpanas T. (2012). Survey on mining subjective data on the web. *Data Min Knowl Disc* 24(3):478-514. doi:10.1007/s10618-011-0238-6.

La risorsa di Italiano Standard ad alta variabilità linguistica per misurare la peculiarità di un corpus

Giovanni De Gasperis[1], Pasquale Pavone[2], Sergio Bolasco[3]

[1]Università degli studi dell'Aquila & Embeds – giovanni.degasperis@univaq.it
[2]Università degli studi di Modena e Reggio Emilia – pasquale.pavone@unimore.it
[3]Università di Roma La Sapienza – sergio.bolasco@uniroma1.it

Abstract
In the automatic analysis of texts, the added value due to the availability of statistical-linguistic resources is indisputable, both for the grammatical tagging of the forms of a corpus, and for the extraction of contents according to their over / under use with respect to the occurrences of a frequency lexicon for identifying the peculiar language. To this end, a corpus is built that is able to estimate the frequency of the so-called Italian Standard as a set of various linguistic typologies. This resource, usable in the TaLTaC software, is of such size as to lend itself to multiple use, both as a whole and in its individual types, each measurable in itself. The first part of the work describes the composition of the lexicon obtained from the corpus. In the second, the resource is tested with respect to a collection of tweets on Russia's war in Ukraine, measuring its specific thematic peculiarity.

Keywords: Standard Italian; frequency lexicons; peculiarity; war in Ukraine

Riassunto
Nell'analisi automatica dei testi è indiscutibile il valore aggiunto dovuto alla disponibilità di risorse statistico-linguistiche, sia per il tagging grammaticale delle forme di un corpus, sia per l'estrazione di contenuti in funzione del loro sovra/sotto uso rispetto alle occorrenze di un lessico di frequenza per l'individuazione del linguaggio peculiare. A tal fine si costruisce un corpus in grado di stimare la frequenza del cosiddetto Italiano Standard come insieme di varie tipologie linguistiche. Questa risorsa, utilizzabile nel software TaLTaC, è di dimensioni tali da prestarsi a un utilizzo plurimo, sia nel suo insieme, sia nelle sue singole tipologie, ciascuna misurabile di per sé. Nella prima parte del lavoro si descrive la composizione del lessico ottenuto dal corpus. Nella seconda si sperimenta la risorsa rispetto a una raccolta di tweets sulla guerra della Russia in Ucraina, misurandone la peculiarità tematica specifica.

Parole chiave: Italiano Standard; lessici di frequenza; peculiarità; guerra in Ucraina

1. Introduzione

Le caratteristiche della lingua comune vs speciale, o le differenze fra linguaggi specialistici vs settoriali (Gualdo, 2021) contraddistinguono molti studi dei linguisti negli ultimi decenni. La discussione rimanda spesso agli studi di sociolinguistica di Berruto (1987) sulle dimensioni diafasica e diamesica, che si incrociano nel continuum del linguaggio (ibid.: 21). Infatti, l'asse diamesico che va dal parlato allo scritto - e più in particolare dal parlato-parlato (dialoghi), e via via al parlato-scritto (interviste), allo scritto-parlato (discorsi in pubblico), allo scritto-scritto (letteratura) si incrocia con l'asse diafasico che va dall'italiano-informale (gergale o trascurato) via via fino allo italiano-formale (burocratico, tecnico-scientifico, aulico). Questi livelli sono espressioni dell'ampia variabilità della lingua, come insieme di linguaggio "comune" e linguaggi "specialistici" (settoriali e/o tecnico-scientifici). Ricostruirne un campione significativo - ad alta "bio-diversità" linguistica - può essere molto utile, dal punto di vista metodologico, come risorsa esterna per i processi di analisi automatica dei testi. L'idea alla base di questo lavoro è quella di costruire una risorsa statistico-linguistica[1] - che potremmo convenzionalmente definire di Italiano Standard (d'ora in poi istd) - di dimensioni tali da prestarsi a un utilizzo plurimo: sia come insieme delle suddette tipologie, sia come sub-lessico specifico di una o più di queste, ciascuna misurabile di per sé. Ciò è possibile in quanto la raccolta delle singole fonti è sufficientemente ampia per esprimere una stabilità della frequenza delle parole considerate, come si può osservare in tabella 1. La risorsa si applicherà, nel software TaLTaC, come lessico di frequenza su un corpus di tweets, al fine di estrarne la peculiarità specifica dei suoi contenuti, a seconda della tipologia prescelta.

2. Le fonti utilizzate per costruire la risorsa

La raccolta di testi compresi nel corpus, che chiameremo "abcd", composto per costruire la risorsa di istd è riconducibile a 8 diversi "Generi" della lingua, che qui di seguito, sommariamente, descriviamo secondo le loro 112 fonti utilizzate: A1 - Parlato: varie tipologie di lingua parlata (corpus del lessico lip di De Mauro *et al*., 1993); corpus per il nuovo vocabolario di base (nvdb) di De Mauro[2]; trasmissioni radio e Tv; focus groups; copioni di film, teatro e serie tv; interviste di vario tipo su argomenti quali: giovani, medicina, sanità, psicologia, scuola, cittadini, mass media; interviste pubblicate sui giornali a personaggi e celebrità;

[1] Per riferimenti ad archivi, corpora e lessici di frequenza vedi le fonti disponibili all'Accademia della Crusca in *https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228*; per un esempio di risorsa multilingue, "illimitata" e perennemente attualizzata cfr. Navigli e Ponzetto, 2012.

[2] Chiari e De Mauro, 2010, p. 29-31.

B2 - Tweets: una metà su temi vari, l'altra metà su tre argomenti: Unione Europea, petrolio in Basilicata, violenza alle donne;
B3 - Comunicazione Mediata dal Computer: e-mail, pagine Facebook, news group, forum di insegnanti, reclami di cittadini, annunci di lavoro, sms da Repubblica;
C4 - Narrativa: letteratura italiana, letteratura straniera, 100 romanzi vincitori del Premio Strega, raccolte di poesie, diari di viaggio, storie di vita, racconti per ragazzi di Rodari, componimenti di alunni delle elementari;
D5 - Stampa: articoli di Repubblica (annata 1988, stralcio delle annate utilizzate nel lessico rep90[3], articoli degli anni 2020 e 2021, fascicoli Robinson di cultura), La Stampa, Corriere della Sera, Sole 24 Ore, Le Monde Diplomatique, L'Osservatore Romano; articoli su tematiche specifiche (lavoro, immigrazione, elettrosmog, Libia, Covid);
E6 - Politica, Economia, Religione: discorsi parlamentari, discorsi programmatici di governo, lessico economico-finanziario (lef[4]), discorsi di Confindustria, lessico bancario, cronache di giornate borsistiche, encicliche papali, omelie;
E7 - Linguaggio Tecnico-Scientifico: testi di linguistica, statistica, informatica, didattica, geografia, enogastronomia e turismo;
E8 – Linguaggio giuridico, Report e Documenti: sentenze delle Corti di Appello, dei Conti e della Cassazione; delibere dell'agcm; Costituzioni Italiana ed Europea, Diritti dell'uomo; relazioni annuali del cnr, rapporti del censis, isfol e inea; bandi di gara di appalti, cartelle cliniche.
Gli 8 Generi, declinati nelle 112 fonti, nel loro insieme rappresentano la suddetta "bio-diversità" e possono raggrupparsi in 5 Tipi: A) Parlato, B) Web, C) Narrativa, D) Stampa, E) Linguaggi specialistici (politico, tecnico-scientifico, giuridico), come evidenziato dalla lettera che precede l'enumerazione dei Generi. Per la loro consistenza si veda la Tab. 1.
Il corpus "abcd", risultato della raccolta dei testi, produce un file di 911 mb in codifica utf8, formato da 60.000 frammenti (porzioni di testo), per un totale di 147,4 milioni di occorrenze, con un vocabolario di 1.535.060 forme grafiche.

3. La costruzione di risorse statistico-linguistiche dell'Italiano Standard

L'insieme di parole appartenenti al nostro lessico di Italiano Standard (istd_Forme) è limitato alle forme grafiche di ogni categoria grammaticale con almeno 5 occorrenze, previa una pulizia del vocabolario di base prodotto dal parsing del corpus "abcd" secondo i seguenti criteri: eliminazione di forme

[3] Bolasco, 2013, p. 261-264.
[4] Canzonetti, 2004, p. 337-350.

grafiche di numeri in cifre, di parole straniere (eccetto quelle correnti nella lingua italiana e presenti nella risorsa linguistica di TaLTaC, ad es. week, woman), di forme grafiche più lunghe di 25 caratteri (ad esempio, parole strillate nel web: NOOOOOOOOOOOOO… oppure stringhe di caratteri qualsiasi disposti in sequenza: ARCpwdse823GEzxFbSOgC3kwceQtL…). Al contrario, sono comprese nella risorsa alcune parole con grafia errata, poiché riflettono errori assai frequenti (usuali nel web o nel parlato trascritto) come le parole terminanti con accento (perchè, ventitrè, quest'ultima anche nella forma ventitre). Infine il lessico istd_Forme contiene 600 parole senza categoria grammaticale[5] dovute a espressioni inusuali (vossignoria, contessina), parole tecniche dialettali (tajarin, bonet), forme arcaiche o poetiche (potea, facea), forme tronche (tien), parole con trattino (socio-professionale, salva-Italia, tam-tam). Un'altra risorsa di Italiano Standard, derivabile dalla precedente, è relativa ai verbi espressi come lemmi. Infatti, data l'elevata flessionabilità di un verbo (44 forme flesse in media per l'italiano), per considerare il suo peso in occorrenze è opportuno cumulare le sue flessioni con almeno 2 occorrenze. Faranno parte di questa risorsa - denominata istd_Verbi - i lemmi verbali con almeno 5 occorrenze[6]. Sempre in Tab. 1 è riportata la consistenza quantitativa totale e delle sue parti.

3.1. Estensione della risorsa ISTD_Forme alle Named Entities

Seguendo i principi dell'Italiano dell'Uso di De Mauro[7] sono state incluse nel lessico istd_Forme alcune entità nominali (Named Entities), anche "multiwords", quali: Nomi propri [Giovanni, Maria Teresa], Toponimi [Italia, New York], Celebrità [Draghi, Putin], Sigle e Acronimi [UE, Covid], Società [Alitalia, Buitoni], Giornali [Repubblica, Stampa]). Le Named Entities sono identificabili sia in fase di normalizzazione in TaLTaC[8], sia grazie all'alta frequenza nel corpus "abcd" di forme grafiche con iniziale Maiuscola. Questa scelta ha prodotto l'inserimento di oltre 20.000 termini con almeno 10 occorrenze. Peraltro, sono state escluse le forme con iniziale Maiuscola di parole comuni identificate grammaticalmente, in quanto hanno una quantità di occorrenze trascurabile rispetto alla stessa forma minuscola (generalmente compresa tra l'1% e il 5% di quest'ultime).

[5] L'assenza di categoria grammaticale è dovuta al fatto che tali forme non esistono nel dizionario di TaLTaC, utilizzato per il tagging grammaticale.

[6] Si noti che considerato un lemma verbale, ad es.<*capire*>, le sue flessioni in *ISTD_Forme* sono 51 (fino a 5 occ.), mentre nella risorsa *ISTD_Verbi* sono 63 perché concorrono al lemma anche le flessioni a soglia 2 occ.

[7] De Mauro (1999), Vol. I, Introduzione § 25 e § 30, pp. XXXVII, XXXIX, XL.

[8] Grazie al dizionario CUCS (Bolasco, 2013, p. 79-81).

Tabella 1. Consistenza di occorrenze - secondo gli 8 Generi e i 5 Tipi - del Corpus ABCD e delle Risorse ISTD_Forme e ISTD_Verbi

		CORPUS "*ABCD*"		RISORSE "*ISTD*"			
		FORME GRAFICHE totali *		FORME GRAFICHE totali **		LEMMI dei VERBI ***	
GENERE	TIPO	GENERE	TIPO (mln)	GENERE	TIPO (mln)	GENERE	TIPO (mln)
1_Parlato	A_Parlato	11.472.228	11,5	10.592.711	10,6	1.682.931	1,7
2_Tweets	B_Web	4.577.516	15,5	3.276.375	12,1	400.748	1,5
3 - Comunicazione Mediata dal Computer		10.962.014		8.882.641		1.102.855	
4_Narrativa	C_Narrativa	37.409.101	37,4	34.660.343	34,7	5.031.991	5,0
5_Stampa	D_Stampa	50.537.610	50,5	46.148.136	46,1	5.151.021	5,1
6_Linguaggio Politico & Economico & Ecclesiale	E_Linguaggi Specialistici	14.686.818	32,4	14.027.042	29,9	1.536.422	2,7
7_Linguaggio Tecnico-Scientifico		8.528.383		7.736.022		637.281	
8_Linguaggio Giuridico & Reports & Documenti		9.192.848		8.085.811		575.203	
		147.366.518		133.409.081		14.435.521	
		** valori dovuti a 1.535.060 forme totali*		*** valori dovuti a 175.980 forme con almeno 5 occ*		**** valori dovuti a 5.512 lemmi con almeno 5 occ*	

3.2. Il continuum delle diverse tipologie linguistiche

Una misurazione dell'intreccio dimensionale diam/diaf del lessico istd_Forme si può ottenere considerando una matrice di dati con in riga i primi 47mila records della risorsa[9] e in colonna le loro sub-occorrenze per Generi e Tipi. Applicando un'analisi in componenti principali (acp) si ottiene il risultato illustrato nel Grafico 1, in cui sulla seconda componente principale si osserva l'ordinamento A-B-C-D-E dei 5 Tipi, fedelmente descritti dai loro Generi, in un continuum[10] dal linguaggio più informale (tweets) a quello più formale (linguaggio giuridico). L'intreccio 3_cmc con 4_narr è giustificato dalla presenza in cmc di parlato-scritto. Il risultato d'insieme dell'acp prova la coerenza della raccolta dei testi del corpus.

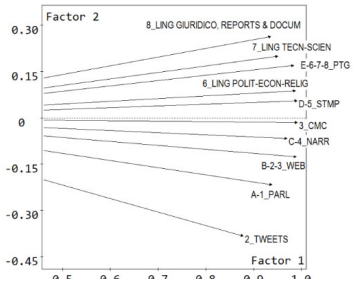


Figura 1. Rappresentazione sul piano fattoriale degli 8 Generi e dei 5 Tipi secondo l'articolazione diafasico-diamesica del continuum della lingua

[9] Le parole con almeno 100 occorrenze, pari alle fasce della legge di Zipf [Alte + Medie + primi decili delle Basse frequenze], (Bolasco, 2013, p. 209-211) assicurano nel nostro caso una copertura del 88% del corpus.
[10] Berruto (2005), p. 127-132.

4. Applicazione

Uno fra i principali obiettivi su come utilizzare le due risorse dell'ISTD è lo studio della pecularità[11] del linguaggio di un corpus sottoposto ad analisi. Questo avviene calcolando uno scarto standardizzato fra le occorrenze normalizzate del corpus e quelle della risorsa, assunta a modello. Nella fattispecie s'intende qui studiare il linguaggio di un corpus di tweets raccolti nei primi giorni di guerra della Russia in Ucraina (24-28 febbraio 2022). L'estrazione dei tweets è stata fatta attraverso le seguenti parole chiave: guerra-Ucraina, Nato-Russia, civili, invasione e profughi, popolazione e armi per un totale di 61.730 messaggi (solo "post"). Il corpus, TW_Guerra, di 11MB, analizzato con TaLTaC 4.0, è stato sottoposto alla risorsa ISTD_forme con l'obiettivo di misurare la peculiarità dei tweets rispetto a due tipologie: "Narrativa" e "Stampa", evidenziandone contenuti differenti. Per ottimizzare la descrizione del risultato, per ciascuna tipologia si riportano i contenuti estratti, citando le radici delle parole selezionate, dai valori più elevati dello scarto standardizzato via via a decrescere. Questa scelta (a parità di spazio) massimizza i contenuti, poiché si cumulano le varianti del sing/plur, femm/masc e delle flessioni verbali. Naturalmente ai primi posti di queste graduatorie rispetto alla risorsa totale o in qualunque sua tipologia - al di là dei nomi dei protagonisti Ucraina, Russia, Putin, Zelensky, Mosca, Kiev, Donbass, Biden, Macron, Cremlino - troviamo il lessico della guerra, che è scontato:

<ucrain, russ,[12] bombard, missil, arm, sanzion, invas, profug, rifug, milit, civil, pace, nuclear, mondial, conflitt, attacc, esercit, aiut, vittim, propagand, internaz … >.

Al contrario, gli scarti maggiori (sempre decrescenti) rispetto alla Narrativa sono relativi a:

- figure (popol, person, uomin, donne, bambin, gent, cittad, famigl)
- situazioni (casa, piazz, foto, immagin, pandem, solidar, uman, accord, appell, folli, resist, assed, aggress, crimin, difend, nazis, sostegn, negozia, diplomaz, Onu, america, ripud, territor, pazz, colpi, lavor, cuore, preghier, risch, occident, accogl, strad, ritard, escal, violen, pericol, protest, responsab, tavolo, allea, regime, nemic, …)
- percezione della condizione di vita (consegu, caus, ragion, rispost, problem, comunit, democr, interess, neutra, ideal, diffic, amic, sicurezz, …).

Considerando invece la peculiarità rispetto alla Stampa, questa risulta più sensibile a:

- figure (bambin, soldat, innocen)

[11] Bolasco (2013), p. 135-141.

[12] Ad esempio gli scarti decrescenti sono: *ucrain* |*i* 718, |*a* 291, |*o* 167, |*e* 122; *russ* |*i* 93, |*o* 47, |*e* 45, |*a* 38.

- toponimi (UE, Odessa, Bielorussia, Polonia, Crimea, Europa, Nato, Italia, Romania, Afghanistan, Finlandia, Occidente, Slovacchia, …)
- azioni e atteggiamenti (scapp, ripud, accogl, invad, fake, inermi, fuga, aiut, video, media, negozia, ucci, ammazz, crepare, massacr, genocidio, pazzo, dittator, pagliaccio, vergogn, fucili, esplos, tank, truppe, resist, ipocris, emergenz, ...)
- linguaggio sboccato (cazz, incazz, merd, schif, culo, caga, coglion, stronz, freg, fott, puttan, …). Quest'ultimo aspetto comprende circa 300 forme, di cui 104 a soglia 3 per un totale di 3.076 occorrenze.

Per quanto riguarda le forme "originali"[13] un esempio per tutte sono i "derivati" di Putin (6.420 occorrenze e peculiarità 383,16 rispetto al totale della risorsa)[14] quali: #Putin (2.246 occ), #PutinWarCriminal (186), #PutinHitler (113), putinian, filo-putin*, putinist, anti-putinian e altre 150 forme grafiche, aventi tutte il valore convenzionale 999.999 dello scarto "non calcolabile".

5. Aspetti computazionali per la misurazione della peculiarità

Il calcolo dello scarto è stato codificato in TaLTaC 4.0[15] in linguaggio Python puro, facendo a meno dell'utilizzo delle librerie di estensione numeriche, contando esclusivamente sull'utilizzo ottimizzato di indici dinamici implementati tramite dizionari e liste connesse ai valori delle variabili di classificazione dei frammenti. Lo scarto standardizzato è calcolato a valle del tagging grammaticale deterministico, confrontando le occorrenze relative della forma nel corpus TW_Guerra a quelle della risorsa statistico-linguistica istd corrispondente alla tipologia prescelta. Il confronto avviene solo se le forme grafiche del corpus dei tweets hanno la stessa categoria grammaticale espressa nella risorsa linguistica dell'italiano (diztal) o nel dizionario di normalizzazione (cucs) delle Named Entities. In particolare, sono state implementate ulteriori ottimizzazioni riguardanti la normalizzazione tramite cucs e le estrazioni delle MultiWord tramite un efficiente algoritmo di individuazione delle concordanze, usando come pivot la prima forma grafica e determinando le forme grafiche successive tramite gli indici di posizionamento dei token nei frammenti. Quest'ultima scelta ha evitato di eseguire ricerche inefficienti sull'intero vocabolario o l'intero corpus. In questo modo si sono potuti ottenere in tempi ragionevoli i risultati delle analisi riportate nei paragrafi precedenti.

[13] Ossia le forme presenti nel corpus e non nella risorsa, che esprimono il massimo della peculiarità.

[14] La parola *Putin* ha un valore dello scarto dalla **Narrativa** pari a 8.036,13 e dalla **Stampa** pari a 504,27.

[15] Per l'evoluzione di TaLTaC cfr. Bolasco (2010), Bolasco e De Gasperis (2017), e il sito https://www.taltac.com

Riferimenti bibliografici

Berruto G. (1987). *Sociolinguistica dell'italiano contemporaneo*. La Nuova Italia Scientifica.

Berruto G. (2005). *Fondamenti di sociolinguistica*. Ed. Laterza.

Bolasco S. (2010). Taltac 2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi. Led Edizioni Universitarie.

Bolasco S. (2013). *L'analisi automatica dei testi*. Carocci Editore.

Bolasco, S. e Gasperis, G. D. (2017). Taltac 3.0. a multi-level web platform for textual big data in the social sciences. In *Data Science and Social Research* (pp. 97-103). Springer, Cham.

Bolasco S. e Pavone P. (2010). Automatic Dictionary and Rule-Based Systems for Extracting Information from Text. In Palumbo, F. *et al*. (eds.) "*Data Analysis and Classification*" 6th Conference of the Classification and Data Analysis, Springer Publ., pp. 189-198.

Canzonetti A. (2004). *La variabilità dei poliformi nel lessico di frequenza del linguaggo economico-finanziario*. Annali del Dipartimento di Studi Geoeconomici Linguistici Statistici Storici per l'Analisi Regionale 2003-2004. Pàtron Editore.

Chiari I. e De Mauro T. (2010). The new basic vocabulary of Italian: problems and methods. Statistica Applicata – *Italian Journal of Applied Statistics*, vol. 22 (1): 23-37.

De Mauro T., Mancini F., Vedovelli M. and Voghera M. (1993). *Lessico di frequenza dell'italiano parlato*. Etaslibri.

De Mauro T. (1999). *Grande Dizionario Italiano dell'Uso*. UTET, 8 Volumi.

Gualdo R. (2021). *Introduzione ai linguaggi specialistici.* Carocci Editore.

Navigli R. e Ponzetto S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, (193): 217-250.

Topic modeling and sentiment analysis: an analysis of tweets about the game Cyberpunk 2077

Alberto Maria De Mascellis, Maria Gabriella Grassia, Agostino Stavolo

University of Naples Federico II – albertomaria.demascellis@unina.it
mariagrabriella.grassia@unina.it agostino.stavolo@unina.it

Abstract
Consumers have become pivotal in sales and even production processes. Studies on brand-user interactions on social network and *prosuming* practices are significant to understand the relationship between companies and customers in digital environments. This interaction also offers many opportunities to obtain data on online user behavior and to better understand how companies improve their strategies. The analysis we propose on the *Cyberpunk 2077* video game by Polish company CD Projekt Red allowed us to study what Twitter users felt and discussed about the game and how the company tried to improve critical aspects based on players' interactions. We performed a Latent Dirichlet Allocation to identify the main discussion topics on Twitter and detect the most discussed issues to highlight the solutions provided by game updates. Moreover, a Sentiment Analysis was carried out for the semantic orientation of users to understand how they expressed themselves about the game.

Keywords: Latent Dirichlet Allocation; Sentiment Analysis; Topic Modeling

1. Introduction

Consumers are crucial in sales and production processes. User-brand interaction on social networks also offers many opportunities to obtain data on user behaviors. Purchase and review/critics processes are milestones to better understand how to elaborate products based on the most widespread people's opinions. Twitter allows users to discuss and comment on a product or service, giving organizations an effective way to analyze customer perspectives and to adapt in the marketplace. This study aims to analyze the Twitter debate in the first year since the release of the game *Cyberpunk 2077* on December 2021 and how the company tried to interact with critics. For this reason, a Latent Dirichlet Allocation (LDA) was conducted to identify the main topics of discussion on Twitter and how the company addressed them in specific periods of *patching*, online updates aimed to fix certain game aspects. Then, a Sentiment Analysis (SA) was carried out to monitor the impressions the game has made on users over time.

2. LDA and Sentiment Analysis

LDA is then one of the most used probability topic models. It is an unsupervised generative probabilistic model of a corpus that uses a "bag of words" approach (Liangjie *et al*., 2010). The basic idea is that documents are represented as random mixtures over latent topics where each topic is characterized by a distribution over words. The model assumes that each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over several words that share a common Dirichlet prior (Blei *et al*. 2003). One method that can be used along with Latent Dirichlet Allocation is Sentiment Analysis (Dahal *et al*. 2019; Yang *et al*., 2018). SA aims to categorize texts according to their polarity (Cambria *et al*.,2013), detecting different emotional states and revealing behavioral patterns. There are three different levels to select the basic unit of analysis: a text-level, a sentence-level, an aspect-level. The text-level considers each text as a whole for evaluating a positive or negative sentiment. At the sentence-level, segmentation of each text is based on its component sentences, calculating the polarity of each sentence, and measuring their overall orientation. At the aspect-level, the segmentation involves the semantic structure, breaking down the text into 'aspects' (attributes or components) and allocating each one an orientation to positivity or negativity (Misuraca *et al*., 2021).

3. Preliminary results

3.1. Data Retrieving and Preprocessing

We extracted statuses from Twitter using the academy access API. We defined the textual corpus using the #CyberPunk2077 hashtag published from December 10, 2020 - when the game was released - to December 10, 2021 - one month after the release of the game's latest update patch. Once removed duplicates by considering the ID of the tweets, the total number of texts was 487.458. Twitter data are free texts and highly unstructured, so they need to be pre-processed with cleansing steps to generate structured data. The preprocessing phases were so structured:

- The text was normalized converting all the letters of the documents into the lower case;
- Special characters, punctuations, numbers, extra white spaces and URL were removed;
- Stopwords, i.e., words with no analytic value, such as "and", "the", etc. were removed;
- Hashtags symbols (#) are removed, retaining the contents;
- Stemming document, to remove suffixes from words and to convert the words into their base form; for example, stemming "played", "plays", and "playing" into "play".

3.2. LDA

There are several automatic methods for LDA to identify the proper number of topics (*k*) contained in a set of documents, most based on computing similarities (or distances) between topics taken in couples over reiterate instances. We relied upon these metrics:

1. Griffith2004: this Bayesian method is based on the computation, through Gibbs sampling, of the posterior probability of a set of models according to all the words of the corpus, namely the likelihood Pr (w | K). The higher the value, the greater the likelihood of the number of topics (Griffiths and Steyvers, 2004).
2. CaoJuan2009: the technique selects the distance between topics, calculated by means of the average similarity of the cosine (*r*), which allows to define the density of a topic. The best value of K is reached when *r* between topics is at a minimum (Cao *et al*. 2009).
3. Arun2010: this measure calculates the symmetric Kullback-Leibler divergence between the distributions of the singular values of the Topic-Word matrix and of the Document-Topic matrix. The best number of topics is identified when the value of the divergence between these two measures in each topic is at its minimum (Arun *et al*. 2010).
4. Deveaud2014: the proposed metric employs the Jensen-Shannon divergence to determine that value of K that maximizes the distance between each pair of topics (Deveaud *et al*. 2014).

Hence, we chose the shortest distance available between metrics, as a reverse ratio of dissimilarity (Holligher, 2018). The number of topics we determined that both minimize Arun and Cao Juan and maximize Griffiths and Deveaud is $k = 8$ as in Fig.1.

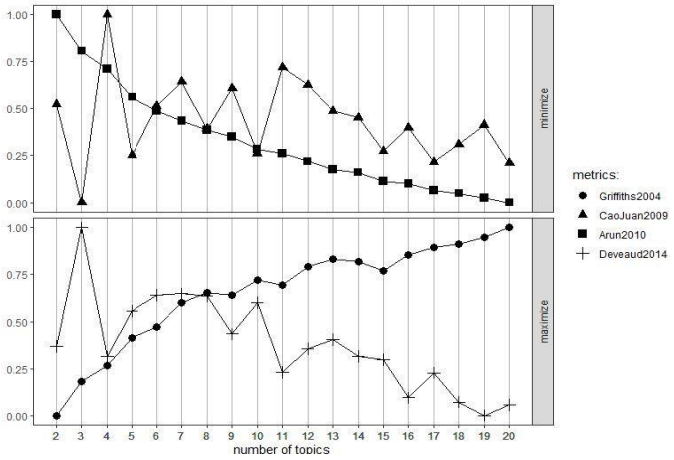


Figure 1. Selection number of k topics

The 10 most frequent words in each topic are shown in Tab. 1:

Table 1. Top 10 frequent words for 8 topics

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6	TOPIC 7	TOPIC 8
Play	Ps5share	V	Photomode	Stream	Release	Check	City
Enjoy	Xboxshare	Enjoy	Virtualphotography	Live	Enjoy	Giveaway	Night
Bug	Ps4share	Art	VGPunite	Twitch	Bug	New	Shutterpunk
Hour	Enjoy	Love	CDProjektred	Play	People	Win	Photomode
Time	Bug	Johnny	Gamergram	Twitchstreamer	Patch	Want	PS4share
Look	Look	Character	Worldofvp	Gamer	Refund	Day	Judy
Story	New	Look	Artistcofsociety	Tonight	Play	Copy	V
Love	Glitch	New	Vpeclipse	Streamer	CDProjektred	Video	PS5share
Best	Gun	Fanart	Developer	Smallstreamer	Year	Play	Xboxshare
Wait	Think	Silverhand	Thecapturedcollective	Twitchaffilate	Xbox	Enter	Nightcity

We have labeled the first topic as "Game experience": the terms "now" and "time" here refer to past periods in the "history" of the game, which is nice to "watch", despite the annoying "bugs" that people "wait" for to be fixed. The second is "Console experience" because of the names of the consoles used but also because of the poor experience reported during playtests on consoles, especially compared to the PC experience. The third topic is "Characters": "V" is the main character, male or female, "Johnny Silverhand" is the main supporting character and in some game dynamics also a "love" interest to be portrayed in "fanart." The fourth topic is about the "Photographic experience": the graphics and the construction of the "world of V" allowed the photographic mode to be used creatively, sharing the results. The fifth topic is about the "Game streaming" practices for the most influential streamers and channels that brought the game "live" during gameplays. In addition, the sixth topic is as fundamental as "Purchase" experience: as we mentioned, the game has been plagued by errors and constant crashes, hence the wait for updates and patches, or in the most extreme cases even a financial "refund." In the seventh topic, there are the economic aspects that we have identified as "Gift users": people who share not only the desire to have a "copy" or "edition" of the game, but who could also win them through a special "giveaway." The last one is about the "World", the "Night city" in which the main story takes place and where "V" and "Judy" reside and act, filtered by the PlayStation's photo mode to remind us of the importance of the graphical and image aspects of the game.

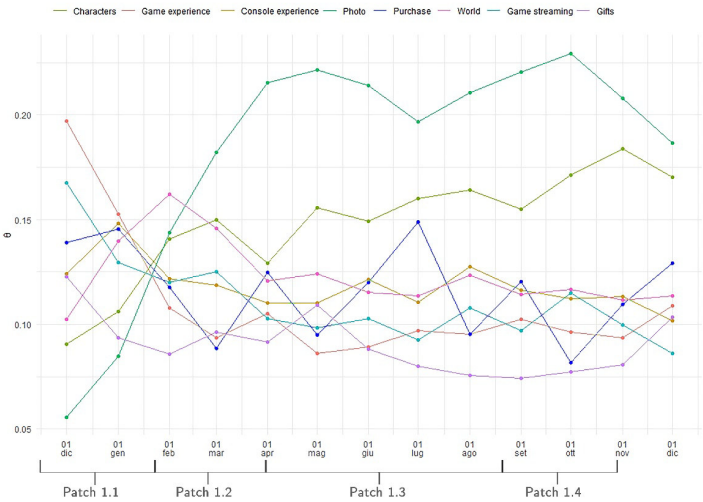


Figure 2. Topic proportions overtime on a monthly interval

Fig. 2 shows the evolution of topics during the four patching periods[1]. Since the company made public what the patches changed[2], we interpreted the topics variations through these modifications too. This relationship is particularly relevant considering aspects like "Purchase" dimension: it had spikes in attention mostly around the corrective patching periods. Also, first patches' initial improvements had reflections on users' perception: it affected "Quests" (game missions) and people appreciated "Characters" topic, that are in fact the "agents" of every quest; it was also based on "Open World" and people dedicated to "Photo mode" to portrait it; "Console experience" also was addressed and purchasers noticed that, but too late to seriously arouse interest over time. This steering revealed its weaknesses when, during the second patching period, the company tried to address the "Gameplay" and "Quests", with numerous voices in the changelog, but confronted with so many problems (testified by the game's withdrawn from the PlayStation Store) that while the "Characters" topic continued to gain interest, "Game experience" and "Game streaming" continued to decline despite the other patches. In the "Purchase" topic emerges the relation of CD Projekt Red with their publisher, Sony, since the selling aspect is joint with the "refund" one in continuous spikes of attention: Sony, in fact, granted refunds until June 2021, where the maximum topic's peak is indeed.

[1] We intended for "patching period" the time span of company's work ending with the patch release

[2] Retrieved https://www.cyberpunk.net/en/news/37373/patch-1-1 on 06/05/2022

3.3. Sentiment analysis

Due to the 280 characters' limitation on tweets, classifying the sentiment of Twitter messages must fall in the previously mentioned sentence-level analysis category (Yu *et al*. 2003). Given a tweet, the *k*-th sentence is as a sequence of its p_k terms. Each term w_{ikj} in the k-th sentence is compared with a dictionary[3] of polarized terms, assigning a score PSwikj of -1 to negative terms, +1 to positive terms and 0 to neutral terms that not listed in the dictionary. The polarity of each term is weighted considering *negator*, *amplifier* and *de-amplifier* terms. This weighting scheme leads to a more effective measure of polarity, as it allows to emphasize or dampen the polarity score of each term. So, the PSsik polarity score of each sentence is calculated as the sum of its weighted term scores PSwikj, on the square-root of the sentence length:

$$PSs_{ik=} \frac{\sum_{j=1}^{pk} PS_{wikj}}{\sqrt{P_k}} \tag{1}$$

Fig.3 shows the sentiment calculated on the sentences and distributed over time. The progression allowed us to contextualize the analysis performed during the LDA and gave us more acknowledgment on the patching periods' effects, as we discuss in the conclusions.

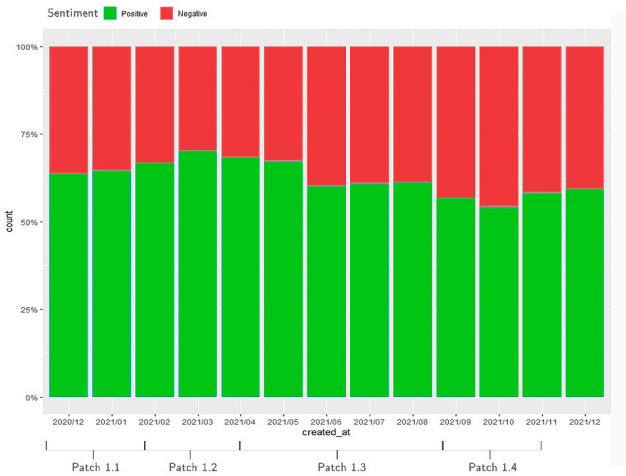


Figure 3. Sentiment proportions (percent) overtime on a monthly interval

[3] For Sentiment Analysis we used *sentimentr* package on R. The package uses the lexicon package's dictionaries.

4. Conclusions

The preliminary results allow to highlight users' reactions following the release of the game and the corrective patches. In this context, it is paramount for game and media producers to know which aspects of their products are relevant to their target consumers and the correct timing for adjustments. The feedback can be directly obtained from users even during the very first life cycles of the games. Analyzing topics extracted through LDA and sentiment expressed by consumers on Twitter, companies would be able to make improvements in their reactions to critical issues in current and future game projects. CD Projekt Red, in fact, has developed several patches to improve and solve bugs in the game's fruition, increasing positive sentiment with the first two patches but then stranding on a too long central patching period that didn't match the expectations, as the trend of the curves on both methods proves. LDA topic modeling and SA are powerful exploratory tools when used on a large collection of Twitter data, but there are many implications: the use of LDA topics calculations' metrics and the choice of the lexicon for SA can affect the results. In future developments will be interesting to merge Sentiment Analysis and Topic Modeling, looking at how sentiment is distributed within each topic.

References

Arun R., Suresh V., Veni Madhavan C. E. and Murthy N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Blei D.M., Ng A.Y. and Jordan M.I. (2003), Latent Dirichlet Allocation. *J Mach Learn* Res 3(Jan)

Cambria E., Schuller B., Xia Y. and Havasi C. (2013), New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28, 15–21.

Cao J., Xia T., Li J., Zhang Y. and Tang S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9), 1775-1781.

Deveaud R., SanJuan E. and Bellot P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1), 61-84.

Dahal B., Kumar S. A. and Li Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1), 1-20.

Griffiths T. L. and Steyvers M. (2004). Finding scientific topics. *Proceedings of the National academy of sciences*, 101, 5228-5235.

Holligher T. (2018). Strategic sourcing via category management: helping air force installation contracting agency eat one piece of the elephant, *Air Force Institute of Technology*, Ohio

Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., li Y. and Zhao L., (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 78, 15169–15211

Liangjie H. and Davison B. D., (2010). Empirical study of topic modeling in Twitter. *In*

Proceedings of the First Workshop on Social Media Analytics (SOMA '10). Association for Computing Machinery, New York, NY, USA, 80–88.
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.
Negara E. S., Triadi D. and Andryani R., (2019). Topic Modeling Twitter Data with Latent Dirichlet Allocation Method, 2019 *International Conference on Electrical Engineering and Computer Science* (ICECOS), pp. 386-390
Ostrowski D.A. (2015). Using Latent Dirichlet Allocation for topic modeling in Twitter, Proceedings of the *2015 IEEE 9th International Conference on Semantic Computing*, pp. 493-497.
Yang S. and Zhang H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering,* 12(7).
Yu H. and Hatzivassiloglou V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proc. Of EMNLP.*

Gli insegnanti dopo l'anno della didattica a distanza

Francesca della Ratta-Rinaldi, Monya Ferritti*

INAPP– f.dellaratta@inapp.org m.ferritti@inapp-org

Abstract
In 2021 INAPP (National Institute for Public Policy Analysis) conducted a web-based survey aimed to secondary school Italian teachers to examine the impact that the new organizational and technological set-up has had on teacher's work, in a year marked by the alternation between face-to-face and distance learning and by numerous uncertainties. 2,152 teachers, scattered throughout the country and representative of the various fields of study, answered to the questionnaire. Five open-ended questions were inserted in the questionnaire. More than half of the teachers provided at least one answer to one of the five open-ended questions, totalling 1,165 respondents. The analysis of textual data shows the most important issues mentioned by the teachers: the new way of teaching, problems related to technology and the management of the pandemic.

Keywords: Text Mining; Taltac2; Iramuteq; Covid-19; Secondary Education; teachers; distance teaching

Riassunto
Nel 2021 l'INAPP (Istituto Nazionale per l'Analisi delle Politiche Pubbliche) ha condotto un'indagine web rivolta agli insegnanti delle scuole secondarie di tutta Italia per approfondire l'impatto che il nuovo assetto organizzativo e tecnologico ha avuto sul lavoro dei docenti, in un anno contrassegnato dall'alternanza tra didattica in presenza e a distanza e da numerose incertezze. Hanno risposto al questionario 2.152 insegnanti, distribuiti su tutto il territorio nazionale e rappresentativi dei diversi indirizzi di studio. Oltre la metà ha fornito almeno una risposta aperta, per un totale di 1.165 rispondenti. L'analisi del corpus ha evidenziato i temi più importanti citati dagli insegnanti: le nuove opportunità didattiche offerte dalla tecnologia, i problemi ad essa connessi e la gestione della pandemia a scuola.

Parole chiave: Text Mining; Taltac2; Iramuteq; Covid-19; Scuole secondarie, Istruzione, didattica a distanza

* Il lavoro è frutto di un lavoro comune. Le opinioni espresse sono da riferire alle autrici e non all'Istituto di appartenenza. Il paragrafo 1 è stato redatto da Monya Ferritti, il 2 e il 3 da Francesca della Ratta.

1. L'indagine

Nel 2021 l'INAPP (Istituto Nazionale per l'Analisi delle Politiche Pubbliche) ha condotto un'indagine web rivolta agli insegnanti delle scuole secondarie italiane per approfondire l'impatto del nuovo assetto organizzativo e tecnologico sul lavoro dei docenti, in un anno contrassegnato dall'alternanza tra didattica in presenza e a distanza e da numerose incertezze (Ferritti, 2021). Obiettivo dell'indagine era approfondire la condizione del docente in seguito alle nuove richieste tecnologiche e professionali, con un focus specifico sulle scuole superiori di secondo grado, più a lungo interessate dalla didattica a distanza.
Hanno risposto al questionario 2.152 insegnanti, distribuiti su tutto il territorio nazionale e rappresentativi dei diversi indirizzi di studio. L'analisi che si presenta di seguito è relativa all'analisi di cinque quesiti aperti introdotti nel questionario su: 1) gli accorgimenti utilizzati durante le verifiche per evitare il rischio di plagio o copiatura; 2) le misure adottate dalla scuola per contrastare la dispersione scolastica; 3) le modalità di intervento per prevenire e/o contrastare il cyberbullismo; 4) le modalità adottate per gestire le richieste di famiglie e/o studenti di particolare complessità e criticità; 5) i punti di forza e di debolezza della propria scuola nell'anno scolastico 2020/2021.
Oltre la metà degli intervistati ha fornito almeno una risposta a una delle domande aperte, per un totale di 1.165 docenti. L'analisi del corpus costituito dall'insieme delle domande aperte fornisce informazioni preziose in particolare sul rapporto dei docenti con la tecnologia.

2. I contenuti del corpus

L'insieme delle risposte aperte costituisce un corpus di 36.629 forme[1], con 5.051 parole diverse e un rapporto forme/vocabolario di 13,8, caratteristico di testi di ridotta ampiezza. Tra le parole tema che compaiono nel testo con una frequenza molto elevata troviamo studenti o alunni (499 occorrenze in tutto), incontri o colloqui (nel complesso 332 occorrenze), famiglie e genitori (319 occorrenze in tutto), debolezza (inteso prevalentemente come punto di, 306 occorrenze), docenti (280 occorrenze), verifiche (226), punti/o di forza (224), classi e classe (218), online (200), controllo (175), organizzazione (148), scuola (147), domande (140), in presenza (135), formazione (113), contatti (97). Si tratta di vocaboli che ci introducono immediatamente al lessico utilizzato dagli intervistati in un anno scolastico particolare (Fig. 1).

[1] L'analisi lessicale (Bolasco, 2013) è stata condotta con TaltaC2 (**Errore. Riferimento a collegamento ipertestuale non valido.**), successivamente è stato utilizzato IRaMuTeQ (http://www.iramuteq.org/) per le procedure di classificazione del testo.

Figura 1. Tag cloud: parole piene con più di 30 occorrenze

Considerata la natura eterogenea del corpus in analisi, composto da risposte a diverse domande, è innanzitutto utile individuare le tematiche (o i profili lessicali) che caratterizzano ciascun gruppo di risposte grazie al calcolo delle parole specifiche per domanda (Tuzzi, 2003). Nella prima domanda[2] gli insegnanti si soffermano sui diversi metodi impiegati: vi sono coloro che utilizzano i sistemi anti-plagio offerti da Google, chi assegna domande con ordine diverso, chi ancora propone compiti che prevedono un ragionamento o ancora chi presenta domande in ordine casuale o sotto forma aperta. Altra questione citata spesso è la richiesta fatta ai ragazzi di lasciare le telecamere e i microfoni accesi, così come quella di spegnere i cellulari.

La seconda e terza domanda sono riferite alle misure adottate dalla scuola per contrastare la dispersione e prevenire il cyberbullismo: si tratta di tematiche per le quali i termini utilizzati sono piuttosto simili. Riguardo al contrasto alla dispersione vi sono riferimenti al monitoraggio continuo, all'importanza dei contatti con le famiglie, alla realizzazione di corsi di recupero, alla dotazione di devices tecnologici per gli studenti che ne hanno fatto richiesta e alla particolare attenzione agli studenti BES. In questa, così come nella terza domanda sul contrasto al cyberbullismo è citato il supporto offerto da équipe di psicologi attraverso la realizzazione di incontri specifici o sportelli di supporto. Per la

[2] Per carenza di spazio non sono pubblicate le tavola. Nel testo sono in corsivo le parole presenti nel testo.

prevenzione del cyberbullismo si fa riferimento anche a conferenze realizzate da esperti esterni (polizia postale) o a specifici corsi di formazione per i docenti, insieme ai percorsi di educazione civica o progetti per sensibilizzare i ragazzi.
Tra gli strumenti utili a garantire la comunicazione con le famiglie, i docenti sottolineano di aver dovuto assicurare una aumentata (massima) disponibilità per far fronte alle richieste di contatto con le famiglie, attraverso telefonate, mail o colloqui personalizzati (online, meet).
Infine, nella domanda sui punti di forza e debolezza sono citati tutti gli ingredienti che possono fare la differenza a scuola, soprattutto in situazioni complesse come quelle affrontate nei due anni di pandemia: l'organizzazione, la dirigenza, la collaborazione tra colleghi, la professionalità dei docenti, la buona o cattiva connessione, la possibilità o impossibilità di utilizzare i laboratori, il problema della eccessiva numerosità degli alunni per classe (classi pollaio), i piani della DDI (didattica digitale integrata), i limiti imposti dalla presenza al 50%, ma anche gli spazi e le tecnologie (Lim, rete, digitale) disponibili a scuola e la gestione dell'attività didattica nel complesso.
Oltre alle parole più frequenti e quelle che caratterizzano i diversi sub-corpus, il contenuto del testo può essere ulteriormente dettagliato grazie all'analisi dei segmenti ripetuti, sequenze di parole che si incontrano nel testo e che rimandano a espressioni di senso compiuto. I segmenti possono essere raggruppati in cinque ambiti citati principali: la didattica, la tecnologia, i riferimenti ai punti di forza e criticità della propria scuola, la gestione della pandemia, e il sistema di relazioni e azioni di supporto messe in campo. Come appena emerso i riferimenti alla didattica sono concentrati sulla questione esplicitamente approfondita nel questionario, vale a dire le verifiche in DAD (didattica a distanza) e il controllo dell'originalità dei compiti. I riferimenti alla tecnologia sono concentrati in particolare sulla necessità di mantenere il contatto visivo con gli alunni attraverso le videocamere e microfoni accesi, sulla buona o cattiva connessione Internet, ma anche sullo scambio di mail o la realizzazione di colloqui online con i genitori, tutti elementi che hanno introdotto cambiamenti notevoli nel modo di insegnare. Nel gruppo di segmenti ricondotti ai punti di forza o criticità, gli insegnanti menzionano innanzitutto le caratteristiche dell'organizzazione della loro scuola (spesso buona o ottima), l'importanza della collaborazione tra colleghi (compattezza corpo docente, coesione tra docenti, ecc.) e della presenza costante della dirigenza e dell'organizzazione amministrativa. Un punto di forza della scuola è anche la professionalità del corpo docente (ottimo team digitale, resilienza corpo docente), mentre tra gli aspetti critici viene menzionato il tema delle classi troppo numerose (classi pollaio), insieme all'aumentato carico di lavoro conseguente allo stravolgimento della vita scolastica. Altre criticità da affrontare sono la mancanza di spazi adeguati, l'eccessiva burocratizzazione, la carenza di organico e la presenza di docenti precari. Vi è infine un gruppo di

segmenti che possono essere ricondotti alle azioni di supporto realizzate dalla scuola e all'importanza delle buone relazioni tra i diversi attori che ruotano intorno alla scuola.
Una volta ricostruite le tematiche principali attorno a cui è costruito il corpus, è interessante fornire alcune indicazioni sul "tono" utilizzato dagli insegnanti nelle risposte, attraverso l'indicatore di criticità che si ottiene grazie al dizionario degli aggettivi positivi e negativi presente in Taltac2. L'indicatore è ottenuto rapportando l'insieme degli aggettivi a connotazione negativa su quelli a connotazione positiva utilizzati in un testo, ed è appunto utile per coglierne il "tono" complessivo. Nell'insieme delle risposte analizzate l'indice è piuttosto basso (26,4%) segnalando una generale positività del testo, che in definitiva riporta una valutazione del lavoro svolto dagli stessi docenti e dalla loro scuola. Le statistiche generali condotte sull'italiano standard mostrano un livello di criticità medio del 40% (Bolasco, della Ratta, 2004), superato di poco soltanto dalle risposte alla domanda sui rapporti con le famiglie (46,3%), che assume quindi l'unica connotazione critica, probabilmente sia per il riferimento alle situazioni di fragilità che rendono necessari contatti più frequenti, sia perché la necessità di assicurare un contatto maggiore con le famiglie è stata vissuta da alcuni come impegno aggiuntivo.
Il livello di positività complessivo del testo è un risultato per certi aspetti inatteso, se si considera il contesto estremamente difficile in cui i docenti si sono trovati a operare, anche tenendo conto delle difficoltà che la didattica a distanza ha comportato per i ragazzi, sia per il loro benessere sia per la riduzione complessiva degli apprendimenti certificata dai test INVALSI. Naturalmente il questionario approfondisce (con i quesiti chiusi che qui non si ha lo spazio di citare) le valutazioni degli insegnanti sul tema della dispersione scolastica e della condizione degli studenti dell'ultimo anno. Tuttavia, in una situazione definita da molti critica per gli adolescenti per via delle conseguenze del lockdown sul benessere psicofisico dei ragazzi (Vicari & Di Vara, 2021), sorprende nell'analisi dei quesiti aperti sia il livello di positività del testo l'assenza di riferimenti sia al disagio degli studenti per l'improvviso stravolgimento della loro routine e sia al tema della mancata formazione dei docenti al nuovo contesto, in cui la DAD ha fatto da detonatore per le "criticità strutturali della scuola italiana" (Ferritti, 2020). A prevalere, nel racconto degli insegnanti sembra soprattutto la notazione dell'eccezionalità del cambiamento tecnologico e organizzativo che la pandemia ha imposto alla didattica.

3. Classificazione del testo: le classi di Iramuteq

Per integrare l'analisi lessicale condotta fin qui, il testo è stato analizzato con Iramuteq che, attraverso la scomposizione del corpus in frammenti (o segmenti di testo) effettua una classificazione gerarchica discendente in cluster

caratterizzati da omogeneità di significato (Reinert, 1983; Ratinaud, 2009). Il software suggerisce una classificazione in quattro gruppi, solo parzialmente riconducibili alle cinque domande originarie. Infatti, mentre l'insieme delle risposte alla prima e ultima domanda sono chiaramente identificabili come cluster distinti, le parole utilizzate per le risposte alle domande sulle misure contro dispersione e cyberbullismo e sul rapporto con le famiglie sono più simili tra loro, e confluiscono nei cluster contigui 2 e 3, come si evince dal dendogramma (Figura 2). Le parole del cluster 4 (in cui confluisce il 29,4% dei segmenti), sono ben distinte dalle altre e ripropongono il tema delle verifiche e dei controlli antiplagio. Il cluster 2 (24,1% dei segmenti) raggruppa tutti i termini riferiti alle misure da adottare per contrastare il cyberbullismo e promuovere attività di ascolto. Il cluster 3 (14,5% dei segmenti) pone il tema della comunicazione con le famiglie (mail, meet, telefonata). Le parole che confluiscono nel cluster 1 (il più numeroso, che rappresenta il 32% dei segmenti) sono più variegate, perché riferite ai punti di forza e debolezza, tra cui spiccano quelle associate alla collaborazione tra i docenti, alla qualità della connessione e tecnologica della scuola, alla disponibilità di spazi, alla dirigenza e all'organizzazione complessiva, elementi determinanti specie in una situazione particolare come quella dell'emergenza pandemica.

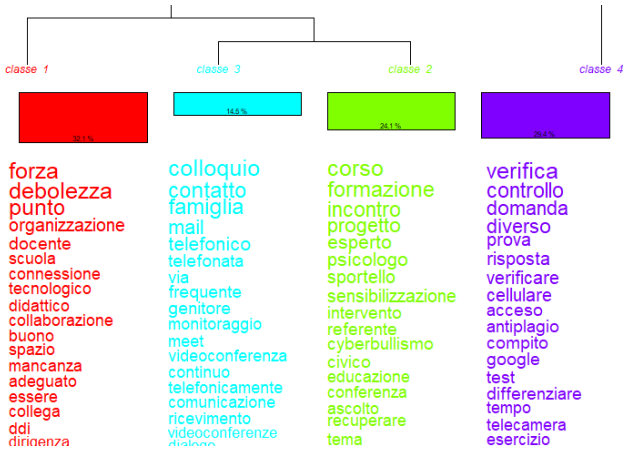


Figura 2. Dendogramma corpus complessivo

La compresenza di valutazioni positive e negative all'interno della stessa domanda ha suggerito di applicare nuovamente la metodologia di classificazione sul solo sub- corpus di risposte sui punti di forza e debolezza della scuola, ottenendo così ulteriori 4 sotto-cluster (Figura 3). Nel sotto-cluster 1, visualizzato sul piano fattoriale in viola e che raggruppa il 21,1% dei segmenti, confluiscono tutte le parole utilizzate per raccontare le difficoltà della didattica, in presenza e

a distanza, nel periodo di pandemia. Se in presenza è difficile stare in classe con la mascherina o tenere le finestre aperte e assicurare comunque un contatto con gli alunni, in DAD i problemi aumentano e raggiungere risultati è più difficile. In particolare, alcuni docenti ricordano l'attenzione che è stata dedicata agli studenti fragili, cui è stato consentito anche nei periodi di chiusura al 100% di frequentare in presenza e di accedere ai laboratori. Più che una valutazione dei punti di forza e debolezza della propria scuola, in questo cluster sembrano confluire le riflessioni degli insegnanti sulle difficoltà insite nella gestione del lavoro didattico nella scuola della pandemia. Nel sotto-cluster 2 (visualizzato in verdino e che raggruppa il 14,7% dei segmenti) confluiscono le parole che caratterizzano l'emergenza pandemica e l'organizzazione necessaria ad affrontarla. In particolare, si fa riferimento alle classi pollaio, che se già costituiscono un problema in sé per la didattica, rappresentano un problema aggiuntivo e irrisolto, per le maggiori difficoltà di distanziamento e prevenzione del contagio che comportano. Altra criticità che confluisce in questo cluster è quella dei mezzi di trasporto, spesso insufficienti a garantire lo spostamento in sicurezza degli studenti. Il terzo sotto-cluster, visualizzato in azzurro e che raggruppa il 27,4% dei segmenti di testo, comprende i termini riferiti all'importanza della coesione e collaborazione di tutte le figure della scuola, in primo luogo docenti e dirigenza, personale Ata e famiglie. Convergono in questo sotto-cluster anche parole come progetto, partecipazione, relazione, rapporto, dialogo e resilienza. Il messaggio sottostante è che, quando questa collaborazione si è verificata, la scuola è riuscita ad affrontare con responsabilità e flessibilità anche la sfida della pandemia. Un punto di forza della scuola è quindi sicuramente una buona squadra, collaborativa, resiliente e coesa. Infine, le parole confluite nel quarto sotto-cluster, visualizzato in rosso e che raggruppa il 36,7% dei segmenti di testo, si riferiscono agli aspetti logistici dell'organizzazione scolastica, sia riguardo le dotazioni tecnologiche (linea, piattaforma, internet) e bontà della connessione (spesso da adeguare), sia in merito alle condizioni generali degli edifici e delle aule e alla disponibilità (spesso carente) di spazi adeguati e di laboratori. In particolare, nella fase di didattica a distanza l'inadeguatezza della rete informatica non ha consentito di approfittare al meglio delle opportunità offerte dalla tecnologia, perché ad esempio insufficiente a ospitare più sessioni in contemporanea.

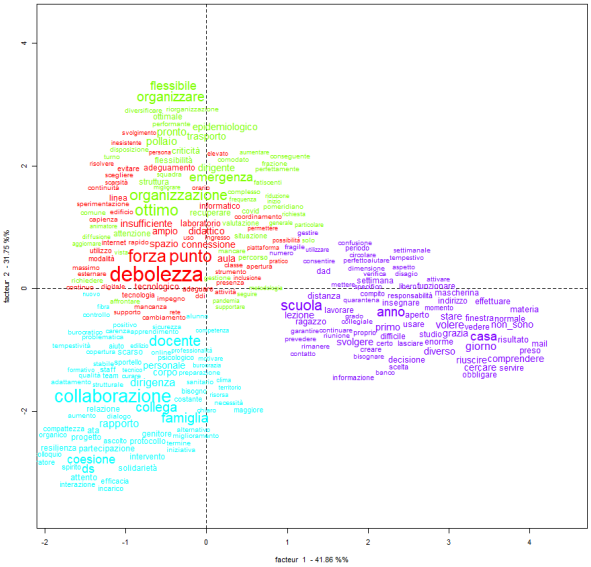


Figura 3. Distribuzione sul piano fattoriale delle sotto classi tematiche individuate all'interno del Cluster 1 (punti di forza e debolezza)

Va da sé che una scuola carente nell'infrastruttura fisica e tecnologica parte svantaggiata nell'affrontare le sfide attuali. È un tema noto, cui molti aspettano risposte anche grazie ai fondi del PNRR e su cui insistono spesso anche gli stessi studenti.

Riferimenti bibliografici

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*, Carocci, Roma.

Bolasco S. e della Ratta-Rinaldi F. (2004). "Experiments on semantic categorisation of texts: analysis of positive and negative dimension", in Purnelle G. *et al* (eds), *Le poids des mots, Actes des 7es Journées internationales d'Analyse Statistique des Données Textuelles*, UCL, Univ. Louvain.

Ferritti M. (2021). "Un anno difficile. L'esperienza del corpo docente delle scuole secondarie di secondo grado nella didattica digitale integrata". In INAPP, *Gender Policies Report* 2021.

Ferritti M. (2020). "Scuole chiuse, classi aperte. Il lavoro degli insegnanti e dei docenti al tempo della didattica a distanza", Sinappsi, X, n. 3.

Ratinaud P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. http://www.iramuteq.org.

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

Vicari S. e Di Vara S. (2021). Bambini, adolescenti e Covid-19. L'impatto emotivo della pandemia dal punto di vista emotivo, psicologico e scolastico. Trento, Erikson.

Evaluating customer satisfaction through Amazon reviews analysis: the Bluetooth earphones example

Alessandro Delmonte, Matteo Farnè

University of Bologna – alessandro.delmonte1997@gmail.com
matteo.farne@unibo.it

Abstract
The present study focuses on how to extract value from textual data related to customer feedback. We analyze a collection of 92,000 Amazon reviews referring to Bluetooth earphones for listening to music, with a twofold aim: accurately classifying reviews according to their sentiment polarity and extracting useful information about the competitive positioning of the considered brands. We employ the Naive Bayes classifier and neural network models, that show a classificatory power between 82% and 85%. We also exploit Semi-Supervised Latent Dirichlet Allocation, whose insights allow, for example, to label Sony products as the most discussed for sound quality, Bose brand as the most criticized regarding the Bluetooth connection, and Apple reviews as the most peculiar for product features and brand ecosystem.

Keywords: Latent Dirichlet Allocation, neural networks, topic analysis, text classification, customer satisfaction, Amazon reviews.

1. Introduction

In today's society, where the use of social media and digital technology permeates many aspects of common life, companies aim at optimizing performance through data collection and analysis. In this respect, value extraction from textual data is an increasingly urgent need (Kushwaha A.K. *et al*., 2021). In fact, the massive use of digital technologies and the habit of sharing opinions on social media generates a huge and disorganized amount of data, difficult to interpret and use in a profitable way. Such data contain potential insights of extreme interest for companies (Raja V.K. *et al*., 2017). Customer Interviews and Focus Groups, that have been among the most used methods for gathering information in the past decades, can now be complemented by the analysis of digital content, produced on a daily base.

For this reason, we conduct a textual analysis based on a real case study: we analyze the Amazon reviews of the main competitors in the sector of "Bluetooth earphones for listening to music". Our statistical results are useful and replicable on other platforms and for any product sector. This research has two objectives:

the correct classification of the analyzed texts according to their sentiment polarity (negative or positive) and the extraction of insights regarding the key aspects of the analyzed products[1]. To reach the first objective, we used the Naive Bayes classifier (Manning, C. D. *et al*., 2008) and neural network models (Rai, B., 2019), while for the extraction of insights we opted for text mining via Semi-Supervised Latent Dirichlet Allocation (Watanabe K., and Zhou Y., 2020).
This paper is structured as follows: in Section 2 we describe the dataset; in Section 3 we describe the methods used to predict the sentiment polarity of the analyzed texts; in Section 4 we present the results regarding the extraction of the topics contained in the reviews. Final considerations follow.

2. Dataset description

For this research, we have chosen to rely on Amazon reviews, as they are available in large quantities, easy to find, and suitable for the use of supervised learning methods, due to the satisfaction rating from 1 to 5 attributed by the buyer. We have collected about 92,000 reviews, referring to 13 different products (7 different brands), all belonging to the sector of Bluetooth earphones for listening to music. Focusing on different competing products allows us to study the sector competition, to investigate performance differences among different brands, and to evaluate the different aspects that influence the choice to purchase. The 13 analyzed products are among Amazon's best sellers in the competitive sector and are summarized in the table below.

Table 1. Brand and Amazon Standard Identification Number (ASIN) of the analyzed products.

BRAND	**ASIN**
Bose	B0748G1QLP
Jabra	B077ZGRVRX - B082Y214W3
Tozo	B07J2Z5DBM - B07RGZ5NKS
Samsung	B07MWCNR3W - B083KVM9VW - B089B658NP
Apple	B07PYLT6DN - B07ZPC9QD4
Sony	B07T81554H - B085VQFZ8Z
Nobrand	B07S93YL9K

The Amazon page of each product can be visited by concatenating the link "www.amazon.com/dp/" with its Amazon Standard Identification Number.
(e. g., www.amazon.com/dp/B07MWCNR3W)

[1] The insights presented in this paper are the result of a research not connected in any way to any company or brand. The analyses are performed in total absence of conflict of interest.

3. Classifying Amazon reviews

The correct classification of Amazon reviews from the provided text is a desirable goal for companies, because this could help distinguishing a satisfied customer from a dissatisfied one, and would lead to optimize the retargeting marketing efforts, thus increasing the Customer Lifetime Value (Jayasanka S.C. *et al*., 2013; Mercanti-Guérin M., 2020). Amazon reviews include a summary judgment of the buyer expressed with the assignment of a score from 1 to 5. In the following analyses the reviews rated from 1 to 3 are considered as negative, those rated 4 and 5 as positive.[2] The methods with the best predictive power are, in our exercise, the Naive Bayes classifier and the classification models based on neural networks.

3.1 Naive Bayes classifier

In order to implement the Naive Bayes classifier, the collected data must be structured as a Document Term Matrix (DTM). In this case, the DTM is a frequency table where the analyzed reviews are in the rows, the words used in the collected texts are in the columns, and each entry shows the absolute frequency of occurrence of a specific word in a given review. The Bayesian classifier is suitable for this context as it uses all the words in the dataset to estimate the polarity of the reviews. The logic behind this classifier is based on the principle of independence, according to which the different words contribute independently to the classification of a review as negative or positive. For each review, the probability of belonging to the class of positive or negative reviews is calculated, and the review is then assigned to the class with larger estimated probability (Manning, C. D. *et al*., 2008). These probabilities are the product of the a priori probability that a review is positive/negative (1) (given by the frequency with which the two classes appear) and the probability that the individual words used appear in the review, given that the review is positive or negative (2) (Manning, C. D. *et al*., 2008).

c): prior probability of a review occurring in class c. N_c: number of reviews in class c. N: total number of reviews. $\hat{P}(t\|c)$: occurring probability of a word t given the class c. T_{ct}: occurrences of word t in training reviews from class c. V: dataset vocabulary. Contains the words used at least once in the analyzed texts.	$\hat{P}(c) = \frac{N_c}{N}$ (1) $\hat{P}(t\|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)}$ (2)

[2] Some studies preferred to exclude 3-point reviews from the analysis (e.g., Haque T.U., et al., 2018). We have chosen to include them in the negative group because they are far below the general average score of the analyzed products (that is 4.4/5). The inclusion of 3-star rated reviews also allowed us to base the analyses of Sections 3 and 4 on the same dataset.

Table 2. The table shows the probabilities estimated by the Naive Bayes classifier of using the indicated words within a review, given that the review is negative or positive.

Words mostly positively connotated			Words mostly negatively connotated		
	P|Negative review	P|Positive review		P|Negative review	P|Positive review
design	0,00131	0,00989	**uncomfort**	0,00892	0,00470
fantast*	0,00293	0,00953	**lose**	0,00951	0,00824
longer	0,00132	0,00929	**reset**	0,00890	0,00232
pocket	0,00118	0,00833	**outsid***	0,00890	0,00100
price	0,00302	0,00619	**quick**	0,00875	0,00194
wireless	0,00275	0,00582	**charger**	0,00862	0,00109
waterproof	0,00332	0,00771	**left**	0,00816	0,00262
cancel	0,00381	0,00637	**fell**	0,00657	0,00291

Table 3. Predictive ability of the Naive Bayes classifier.

TEST SET		**Actual class**		**Predictive Ability**	TRAINING SET		**Actual class**		**Predictive Ability**
		0	1	Accuracy : 82,12%			0	1	Accuracy : 83,92%
Predicted class	0	8222	1920	Sensitivity : 89,84%	**Predicted class**	0	16586	3317	Sensitivity : 91,33%
	1	3566	16982	Specificity : 69,75%		1	6551	34925	Specificity : 71,69%

The predictive power is very good, and it is possible to say that negative and positive reviews are characterized by the use of different words, as well as that this difference is sufficiently captured by the Bayesian classifier. This method is effective but has an important construction bias: treating words as independent does not consider their order or position within a sentence. The neural network models discussed in the next chapter are a possible classification method able to overcome this drawback.

3.2 Neural network models

For the implementation of the neural network models described in this paragraph, we used the "Keras" and "TensorFlow" R packages.

The implementation requires a data preparation phase in three steps (Rai, 2019):

- tokenization: operation that partitions the text into smaller units that can be processed. Sentences are broken down into the single words they contain;
- conversion of text into integers: the occurrence frequency of each word is counted (in all the collected reviews) and these are sorted and numbered in descending order of frequency. The collected reviews are then rewritten by replacing the words with the relative integers that indicate their position by frequency of use;
- padding and truncation: in order to be processed, the reviews must all have the same length. A number of words common to all reviews must therefore

be chosen and these must be adjusted, by padding or truncation, to the designated length.

Table 4. The following table summarizes four tested neural network models, reporting the key parameters for their construction and their predictive ability. The first, Simple Neural Network (SNN), is a feedforward network that uses ReLu (Rectified Linear Unit) as activation function. Models two, three and four are Long Short-Term Memory, a special type of Recurrent Neural Networks that makes possible to capture long-term dependencies between words (Rai, B., 2019).

id	different words used	sentence length	neurons number in 1st layer	neurons number in 2nd layer	layer type	epochs	batch size	optimizer	parameters estimated per word	Accuracy (training set)	Accuracy (test set)
1) SNN	800	31	32	-----	Simple	3	36	rmsprop	12	85,50%	83,99%
2) LSTM	800	31	32	-----	LSTM	7	36	adam	13	86,27%	84,65%
3) LSTM	800	31	32	32	LSTM	8	36	adam	13	87,06%	84,54%
4) BID. LSTM	800	31	32	-----	BID. LSTM	6	36	adam	13	86,24%	84,42%

All tested neural network models use binary cross-entropy as loss function and have very good predictive power. In this specific context, the best is the Long Short-Term Memory (LSTM) model, a recurrent network with a single hidden layer. The non-recurrent Simple Neural Network (SNN) model is probably too simple, which prevents it from finding more meaningful patterns of words that help the correct classification (Zhou Z., Xu L., 2016). Neural networks of type 3) and 4) are more complex models. However, the second hidden layer of model 3) does not in fact increase the predictive ability. Similarly, the peculiarity of the Bidirectional Long Short-Term Memory (BID. LSTM) model, that is to search for patterns also in the opposite direction to the reading one, does not seem to significantly help.

4. Insights Extraction by Semi-Supervised Latent Dirichlet Allocation

For the extraction of insights, we have chosen to use the Semi-Supervised Latent Dirichlet Allocation (SLDA), a technique that can identify topics (intended as collection of words structurally interlinked) according to a context-personalized dictionary that is provided as input (Watanabe K., and Zhou Y., 2020). SLDA is implemented using the R packages “quanteda” and “seededlda”. This methodology is tested with two different dictionaries built with two different logics.

Table 5. The table shows the single words that specify the topics of the two dictionaries used for the Semi-Supervised Latent Dirichlet Allocation.

DICTIONARY 1	DICTIONARY 2
TOPIC1 **sound** = *("sound*", "bass", "voice", "volum*", "calls", "music")*	TOPIC1 **soundpositive** = *("sound*", "excellent", "incredible", "superb", "deep", "superior", "impressive", "surprised", "outstanding")*
TOPIC2 **price** = *("price*", "value*", "conven*", "money")*	TOPIC2 **soundnegative** = *("tinny*", "noise*", "decent*", "ok", "however", "poor", "muffled", "okay", "average", "cons")*
TOPIC3 **connection** = *("connec*", "bluet*", "app*","connection", "wireless", "disconnect", "disconnecting")*	TOPIC3 **price** = *("price*", "value*", "conven*", "money")*
TOPIC4 **charge** = *("charge*", "charging*", "battery","charger")*	TOPIC4 **connectionpositive** = *("connec*", "connection", "easy")*
TOPIC5 **fit** = *("siz*", "stay", "fall", "little", "big", "comfor*")*	TOPIC5 **connectionnegative** = *("connec*", "connection", "lose", "disconnect", "constantly", "reset", "doesnt", "wouldnt", "losing", "frequently")*
TOPIC6 **problem** = *("one", "noise", "issue", "issues", "problem", "stopped", "broke", "break", "broken", "breaks", "miss", "missing", "never")*	TOPIC6 **chargepositive** = *("charge*", "charging*", "battery","charger", "long", "longer", "easy", "wish", "nice", "enough", "entire", "whole", "take")*
TOPIC7 **positive** = *("great", "good", "love", "best", "amazing")*	TOPIC7 **chargenegative** = *("charge*", "charging*", "hold", "dead", "stopped","months", "anymore", "sucks", "leave", "minutes")*
TOPIC8 **leftrightone** = *("left", "right", "one")*	TOPIC8 **fit** = *("siz*", "stay", "fall", "little", "big", "comfor*")*
	TOPIC9 **problem** = *("one", "noise", "issue", "issues", "problem", "stopped", "broke", "break", "broken", "breaks", "miss", "missing", "never")*
	TOPIC10 **positive** = *("great", "good", "love", "best", "amazing")*
	TOPIC11 **leftrightone** = *("left", "right", "one")*

The key areas on which the various topics are built have been identified through an initial exploratory analysis[3] and consist of sound, price, connection, recharge, fit, general malfunction problems and the specific malfunction of one of the two earphones. There is also a specific topic containing all the positively connotated words. In the first dictionary, the topics are defined with the aim to detect when a user discusses a key aspect of the product within his/her review. The second dictionary instead tries to understand the sentiment of a buyer while talking about the various features. LDA estimates a membership probability for each review to belong to each specified topic. Words in reviews that cannot be attributed to any topic are assigned to the additional topic "others". By dividing the obtained results by brand and calculating the average of the estimated membership probabilities for each review with respect to each topic, it is possible to construct the graphs in Figure 1.

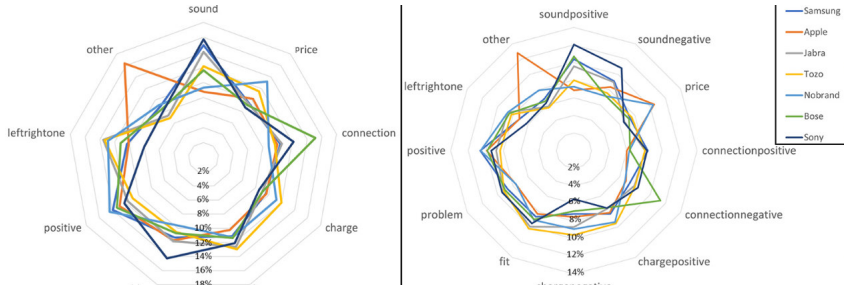


Figure 1. The graphs show the average membership that the SLDA estimates for the reviews of each brand with respect to each topic specified in the two dictionaries. Each circle indicates an increase / decrease of 2 percentage points, compared to neighboring circles, in the average diffusion of the various topics.[4]

[3] The topics have been determined by a detailed preliminary association analysis aiming at identifying the words most frequently appearing together with each of the keywords representing each topic.

[4] The sum of the membership proportions for each review adds up to 1.

The described analysis allows us to note that:

- Sony is the most discussed brand in terms of sound quality.
- "Nobrand" product reviews often contain a reference to price and to the "positive" topic.
- Jabra generally has average characteristics, except for problems with one of the two earphones (topic "leftrightone"), that seems to be frequent.
- Apple reviews are frequently focused on price and the topic "other". The reviews of this brand are expectedly mostly focused on the Apple ecosystem (among the most frequent words in Apple product reviews we find "apple", "iphone", "Airpods", "air", "pro").

Finally, we used the SLDA estimates as regressors in a logit model, in which the dependent variable indicates the true polarity of reviews. This allows us to have an estimate of the impact of each topic on buyer satisfaction.

Table 6. Logit regression model output for Dictionary 1 (left) and Dictionary 2(right), where the dependent variable is the true polarity of reviews.

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.41902    0.07748  -5.408 6.38e-08 ***
sound         0.80639    0.10737   7.510 5.90e-14 ***
price         4.48709    0.12835  34.961  < 2e-16 ***
connection   -1.41537    0.11446 -12.365  < 2e-16 ***
charge        1.11038    0.11473   9.678  < 2e-16 ***
fit          -0.64974    0.10620  -6.118 9.47e-10 ***
problem       0.30198    0.11393   2.651  0.00803 **
positive      8.89587    0.14677  60.613  < 2e-16 ***
leftrightone -4.17298    0.10797 -38.649  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81479  on 61378  degrees of freedom
Residual deviance: 64912  on 61370  degrees of freedom
AIC: 64930
```

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.31051    0.09693  -3.204 0.001357 **
soundpositive        2.23895    0.15360  14.577  < 2e-16 ***
soundnegative       -0.15991    0.14127  -1.132 0.257654
price                4.78061    0.16444  29.072  < 2e-16 ***
connectionpositive   2.31780    0.14763  15.700  < 2e-16 ***
connectionnegative  -4.05859    0.14710 -27.591  < 2e-16 ***
chargepositive       2.71419    0.14834  18.298  < 2e-16 ***
chargenegative      -2.50951    0.13743 -18.260  < 2e-16 ***
fit                 -0.63988    0.13368  -4.787 1.69e-06 ***
Problem              0.50999    0.14596   3.494 0.000476 ***
positive            10.79247    0.19058  56.630  < 2e-16 ***
leftrightone        -5.85586    0.15345 -38.161  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81479  on 61378  degrees of freedom
Residual deviance: 63491  on 61367  degrees of freedom
```

Observing the coefficients estimated for Dictionary 1, we can understand whether a specific topic has a negative or positive impact on customer satisfaction. The topics of the second dictionary are instead already constructed with negative/positive polarity. In this respect, the estimated coefficient signs and significance mostly confirm the appropriateness of the topic labelling. In particular, the coefficients are mostly very significant, and their sign agrees with expectations: Dictionary 2 manages to effectively distinguish the topics with positive connotation from those with negative connotation.

5. Final remarks

In this paper, we tested some methodologies for the classification of Amazon reviews according to their polarity and for the extraction of insights useful for the analysis of the competitive sector of Bluetooth earphones for listening to music. The Naive Bayes classifier and neural network models have a predictive ability between 82% and 85%, that is an excellent result in a real context in which

there is a strong subjective component in the evaluation of a review as positive or negative.
Semi-supervised Latent Dirichlet analysis corroborates the results of the previous exploratory surveys, leading us to a number of insights that turn out to be useful to understand the competitive positioning of different brands and the strengths / weaknesses of each analyzed product.

References

Haque T. U., Saber N. N. and Shah F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE.
Jayasanka R. A. S. C., Madhushani T., Marcus E., Aberathne I. A. A. U. and Premaratne S. (2013). Sentiment analysis for social media. In *Information Technology Research Symposium* (Vol. 11, p. 22).
Kushwaha A. K., Kar A. K. and Dwivedi Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights*, 1(2), 100017.
Manning C.D., Raghavan P. and Schütze H. (2008). Chapter 13 Text classification and Naive Bayes. *An Introduction to Information Retrieval*., 1(6). Cambridge University Press.
Mercanti-Guérin M. (2020). The Improvement of Retargeting by Big Data: a Decision Support that Threatens the Brand Image?. *European Journal of Marketing and Economics*, 3(3), 49-59.
Rai B. (2019). Advanced Deep Learning with R. Packt publishing ltd.
Raja V.K., Sukhwani S. and Khots D. (2017). Enhancing Customer Experience through Text Analysis of Survey Comments.
Watanabe K. and Zhou Y. (2020). Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review*, 0894439320907027.
Zhou Z. and Xu L. (2016). Amazon Food Review Classification using Deep Learning and Recommender System.

Du discours sur le travail avec les chevaux aux valeurs communes des professions anthropoéquines

Vanina Deneux – Le Barh

IFCE - INRAE – vanina.deneux@inrae.fr

Abstract
This contribution aims to understand to what extent the lexicometric analysis of a corpus allows us to reveal the existence of an anthropo-equine professional identity that goes beyond the traditional compartmentalisation of the "horse worlds". A survey was conducted with 108 women and men working with horses in different professional fields. The interviews constitute a corpus that was analysed with the Iramuteq software. The resulting down hierarchical clustering and the chi2 test of association of a variable to classes allow the identification of social representations relating to the living and working relationships between humans and horses. It is thus possible to distinguish between the materiality of each professional group and the commonly shared values.

Keywords: down hierarchical clustering, lexical world, animal work, horse, profession, Iramuteq

Résumé
Cette contribution a pour objectif d'appréhender dans quelle mesure l'analyse lexicométrique d'un corpus permet de mettre au jour l'existence d'une identité professionnelle anthropoéquine qui dépasse le cloisonnement traditionnel des « mondes du cheval ». Une enquête a été menée auprès de 108 femmes et hommes travaillant avec des chevaux dans différents domaines professionnels. Les entretiens constituent un corpus qui a été analysé avec le logiciel Iramuteq. La classification hiérarchique descendante qui en découle et le test d'association du chi2 d'une variable aux classes permettent de repérer les représentations sociales relatives aux relations de vie et de travail entre humains et chevaux. Il peut ainsi être distingué ce qui relève de la matérialité de chaque groupe professionnel et les valeurs communément partagées.

Mots clés : classification hiérarchique descendante, mondes lexicaux, travail animal, cheval, profession, Iramuteq

1. Introduction

Depuis plus de six millénaires, humains et chevaux entretiennent des relations de travail dans des domaines d'activité qui se sont sans cesse renouvelés au cours

de l'histoire (agriculture, guerre, transport, sport, loisirs, art…). À la suite de ce long processus sociohistorique, trois principaux « mondes du cheval » se sont cristallisés en raison de leurs habitus et leurs rapports de classes. Tout d'abord, il y a le monde des courses hippiques marqué par l'aristocratie et la haute bourgeoisie ; vient ensuite celui des sports équestres dominés par les classes moyennes et moyennes supérieures ; enfin la sphère des chevaux dits territoriaux (intervenants dans les villes et en agriculture) associée aux classes ouvrière et paysanne. Il en ressort que chaque monde porte un regard critique et lourd de préjugés à l'encontre des deux autres. Pourtant, quel que soit le type de métier, tous entretiennent des relations de travail avec un même animal : le cheval. Nous pouvons alors nous demander si les dissensions entre ces mondes du cheval persistent dès lors que l'on étudie la quotidienneté des relations de travail que les humains entretiennent avec leurs chevaux. À partir de ce questionnement et dans le cadre d'une recherche doctorale, une enquête de terrain en France métropolitaine a été menée auprès de cent-huit professionnels issus de ces trois grands mondes du cheval. Cette enquête consistait principalement en des entretiens semi-directifs qui intégralement retranscrits ont constitué un corpus de données textuelles analysé avec l'aide du logiciel Iramuteq (Deneux – Le Barh, 2021a). Le but de l'analyse lexicométrique est, ici, de repérer les représentations sociales relatives aux relations de vie et de travail avec les chevaux auxquelles les professionnels se réfèrent communément. Après avoir présenté les caractéristiques sociologiques du corpus, j'analyserai les résultats obtenus à partir de la classification hiérarchique descendante du corpus selon la méthode Reinert. Ainsi, je montrerai en quoi l'analyse lexicométrique a permis de mettre au jour les modalités d'un vivre ensemble interspécifique laissant transparaitre l'existence de valeurs qui dépassent les rapports de classes et les représentations sociales traditionnelles.

2. Le corpus et son analyse

2.1. Présentation du corpus

L'objectif de cette enquête, conduite de février 2017 à avril 2019, était donc d'obtenir une vision panoramique des mondes du cheval afin de rendre compte de sa diversité, mais aussi de repérer l'existence, ou non, d'éléments communs à l'ensemble de ces domaines. Le choix des individus relevait de plusieurs critères dont le sexe, l'âge, la proximité du lien avec le cheval, le métier principal et le domaine professionnel. Les 108 locuteurs se répartissent de la manière suivante : 65 % d'hommes et 35 % de femmes. L'âge est compris entre 19 et 82 ans. 67 % des professionnels ont un lien direct avec leurs chevaux, c'est-à-dire qu'ils vivent et travaillent quotidiennement avec leurs propres équidés ; 18 % ont un lien indirect avec les chevaux, i.e. ils travaillent quotidiennement avec des chevaux, mais qui ne sont pas les leurs, à l'instar des vétérinaires. Enfin, 15

% des interrogés n'ont pas de liens physiques avec les chevaux dans leur travail, par exemple des personnes travaillant dans des institutions. Par commodité d'analyse ces trois mondes ont été divisé en cinq domaines professionnels où se distribuent 19 métiers (figure 1).

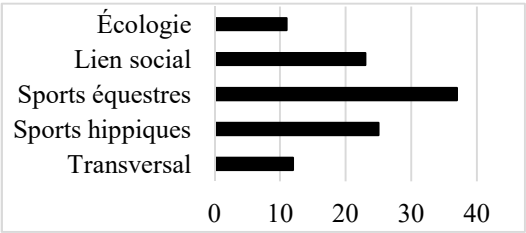


Figure 1. Répartition des enquêtés par domaine.

Le premier domaine professionnel correspond aux chevaux territoriaux (écologie). Il regroupe les fonctions de transport de personnes, la collecte des déchets, les travaux de cantonnier ainsi que le travail en agriculture (maraichage, vigne) et en foresterie. Le deuxième groupe, le lien social, réunit ceux dont le cœur de métier est le spectacle, l'équithérapie, mais aussi la sécurité publique. Le troisième relève des sports équestres. Il rassemble les enseignants d'équitation, les cavaliers professionnels qui sont entendus ici comme des sportifs de haut niveau et les soigneurs. Le domaine suivant est celui des sports hippiques. Il est composé des entraineurs de courses de trot ou de galop, des jockeys et des soigneurs. Le dernier groupe intitulé transversal est le plus hétéroclite puisqu'il réunit des métiers du soin aux chevaux (vétérinaires, maréchaux-ferrants…), des métiers dits connexes, par exemple les selliers ou des individus travaillant dans les institutions de la filière équine. C'est dans cette catégorie que l'on retrouve majoritairement les individus ayant des liens indirects ou qui n'ont pas de liens avec les chevaux.

2.2. Le recours à un logiciel de lexicométrie

Du fait de mon implication personnelle dans le domaine des sports équestres et de la masse de données textuelles recueillies, il m'est apparu opportun de recourir à un logiciel de lexicométrie, car il traite le corpus en réduisant, pour partie, les biais subjectifs propres à la bonne connaissance d'un terrain ainsi que les présupposés du chercheur. Cela permet également de mettre au jour les régularités, les symétries cachées ainsi que les oppositions (Reinert, 1990) des différents discours. Par ailleurs, le recours à ce type d'outil d'analyse est particulièrement pertinent pour les gros corpus, car il permet d'établir des mondes lexicaux stabilisés qui rendent compte des représentations collectives

(Ratinaud, Marchand, 2012). Ainsi, les logiciels de lexicométrie permettent d'objectiver les différentes intuitions qu'un chercheur peut avoir vis-à-vis de son corpus tout en dévoilant certains aspects qu'une analyse manuelle conduite sur un si grand nombre d'entretiens ne relèverait probablement pas (Poudat, Landragin, 2017). Pour ma part, j'ai choisi de recourir au logiciel libre Iramuteq (Ratinaud, 2009) qui a implémenté la classification hiérarchique descendante de la méthode Reinert. Les entretiens semi-directifs ont duré en moyenne une heure, ils ont tous été enregistrés puis intégralement retranscrits. Suite à cela, tous ont été nettoyés des mots phatiques tels que les tics de langage, les hésitations, les répétitions ne prêtant pas à incidence et j'ai retiré la formulation de mes questions ainsi que mes relances. En revanche, j'ai laissé les fautes de prononciation et les néologismes. J'ai ainsi obtenu un corpus textuel de plus de 1100 pages soit plus de 3,5 millions de mots. Pour chaque texte, six variables étaient codées : un numéro correspondant au nom de l'enquêté, le sexe, la tranche d'âge, le type de lien, le secteur (privé ou public), le domaine (illustration 1) et le métier. L'écriture du codage est la suivante : **** *nom_ *sexe_ *age_ *lien_ *secteur_ *dom_ *metier_. Au total, les 108 entretiens forment un corpus de 754 066 occurrences (la moyenne par texte est de 6 982 occurrences), 10 286 formes lemmatisées dont 3 756 hapax (0,50 % des occurrences et 36,52 % des formes).

3. Les vérités partagées par les « mondes du cheval »

3.1. La classification hiérarchique descendante du corpus

L'intérêt de la classification hiérarchique descendante (CHD) élaborée par Reinert est qu'elle permet « de pister [la] démarche identitaire d'un auteur individuel ou collectif, à travers les *mondes lexicaux* » (Reinert, 1993, p.37). En effet, chaque classe d'une CHD est constituée d'une thématique majeure qui « peut être définie comme un ensemble de formes pleines cotextuelles liées entre elles par leur objet et leur contexte » (Ratinaud, Marchand, 2015, p. 58). Lorsqu'il s'agit d'un corpus regroupant plusieurs enquêtés, comme c'est le cas ici, on arrive à une production de « lieux communs » qui « peuvent s'imposer davantage à l'énonciateur qu'ils ne sont choisis par lui, même si celui-ci les reconstruit, leur donne une coloration propre » (Reinert, 1993, p.12). La CHD effectuée sur le corpus des relations de vie et de travail entre humains et chevaux a été initiée avec une demande de 40 classes terminales, le logiciel en a retenu 18 permettant de couvrir 90,55 % des segments classés. Le dendrogramme de la CHD (figure 2) fait état d'une partition initiale des mondes lexicaux. À gauche se regroupent les classes 11, 10, 1, 9, 8, 7, et 17 soit 45,4 % des segments de textes ; à droite, les classes 18, 14, 12, 3, 2, 16, 5, 4, 6, 13 et 15 soit 54,6 % des segments du corpus. Cette démarcation sépare ce qui dans le discours des enquêtés est propre au cheval et à la matérialité de son travail (à gauche) et ce qui relève d'une part des modalités du travail interspécifique ainsi que l'engagement des humains dans

le travail avec les chevaux (à droite). Plus précisément, la classe 1 est constituée des thématiques de l'initiative et de la mobilisation du corps du cheval dans l'activité.

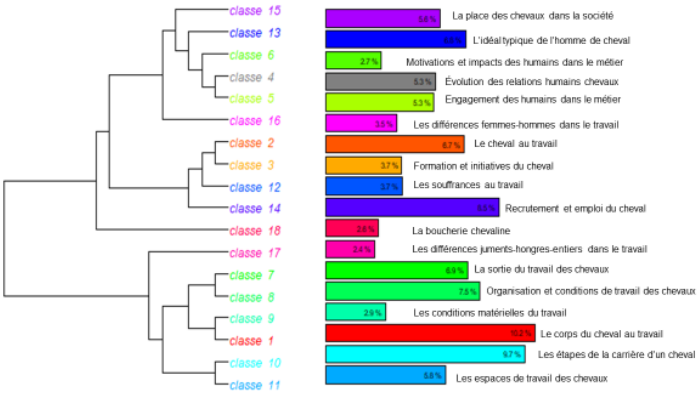


Figure 2. Dendrogramme relatif à la CHD à 18 classes du corpus sur les relations de vie et de travail entre humains et chevaux.

La classe 2 réunit celles du cheval au travail, des affects du cheval au travail, de la collaboration du cheval, du début du travail et de la souffrance des humains et des chevaux au travail. La classe 3 rassemble les initiatives du cheval, sa formation et son activité. La classe 4 regroupe l'évolution des relations anthropoéquines, les mondes du cheval, l'apport du cheval pour le professionnel et l'impact des associations de protection animale sur les métiers anthropoéquins. La classe 5 reprend les thèmes de la relation entre humains et chevaux, l'apport du cheval pour le professionnel, la communication entre humains et chevaux, la signification du travail avec les chevaux pour les professionnels et les caractéristiques de l'homme de cheval.

La classe 6 assemble les thématiques de la famille, de la motivation à travailler avec les chevaux, de la signification du travail avec les chevaux et de l'apport du cheval pour le professionnel. La classe 7 aborde la sortie du travail du cheval, les conditions de vie pour le cheval, ses conditions de travail. La classe 8 regroupe l'organisation du travail avec les chevaux, leurs conditions de vie, le quotidien des professionnels, la famille, les conditions de travail des chevaux, le début du travail et l'intérêt du travail pour le cheval. La classe 9 réunit les conditions de travail et la souffrance au travail du cheval. La classe 10 rassemble les thématiques de la carrière du cheval, de son recrutement à sa sortie du travail, de la motivation des humains à travailler avec les chevaux, de la pérennité du travail avec eux et de l'organisation du travail. La classe 11 réunit les initiatives du cheval, la carrière,

les contraintes du cheval au travail et le recrutement des chevaux. La classe 12 assemble la souffrance au travail des humains et des chevaux, l'apport du cheval pour le professionnel, la souffrance au travail du cheval, le quotidien du professionnel, les conditions de travail et le cheval au travail. La classe 14 regroupe le recrutement des chevaux, leur carrière, leur formation, les conditions de travail et la communication entre humains et chevaux. La classe 15 est relative à l'évolution de la place des chevaux dans la société. La classe 16 a trait aux différences de sexe des humains dans le travail avec les chevaux. À la classe 17, ce sont les différences de sexe des chevaux dans le travail avec les humains qui sont abordées. Enfin, la classe 18 relève de la thématique de l'envoi des chevaux à l'abattoir. En première lecture, ces mondes lexicaux que donne à analyser la CHD pourraient laisser penser que les représentations sociales sont communes à l'ensemble des interlocuteurs quel que soit son domaine professionnel dépassant de facto les clivages des « mondes du cheval ».

3.2. Les relations statistiques entre discours et variables

Comme cela a été expliqué plus haut, chaque enquêté a été associé à sept variables. Ce qui permet d'établir des discriminations statistiques. En effet, l'association d'une variable, ici celle du domaine professionnel, à une classe mesurée par un test du Chi2 permet de montrer si certains discours sont plus liés à un domaine qu'un autre.

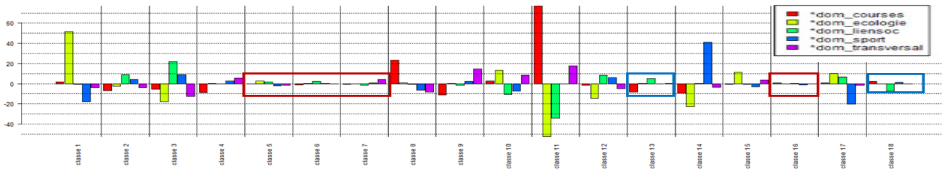


Figure 3. Chi2 par classe de la variable domaine

L'illustration 3 confirme que pour la majorité des classes, il existe bel et bien une variabilité des discours selon le domaine professionnel. Cette variabilité se renforce lorsqu'il s'agit des classes relatives aux spécificités de métiers à l'instar de la classe 11 relative aux espaces de travail tel qu'un hippodrome ou un terrain de concours par exemple. Il en va de même pour la classe 14 qui correspond aux modalités de recrutement des chevaux et de leur emploi : il n'est pas attendu les mêmes critères physiques et mentaux entre un cheval de débardage et un cheval concourant à haut niveau dans une discipline olympique ; de même les corps (classe 1) ne sont pas mobilisés de la même manière. En revanche, il existe bien un ensemble de représentations sociales partagées par l'ensemble des groupes

professionnels interrogés. L'absence de variabilité des classes 5 et 6 indique que, quel que soit le domaine professionnel de provenance, les hommes et les femmes s'engagent dans des métiers anthropoéquins pour les mêmes raisons à savoir la centralité affective qui est propre à tout métier anthropozoologique (Porcher & Estebanez, 2019). L'engagement vocationnel, voire sacrificiel, est très fort et rappelle notamment celui des professions artistiques (Sapiro, 2007). Un autre élément de l'engagement dans une profession anthropoéquine est que son cœur de métier est charnel, c'est du corps à corps. Cela nécessite une exacerbation des sens qui en fait une profession plus sensorielle et intuitive que rationnelle et intellectuelle : le verbe vient après le corps. Malgré des variabilités individuelles, plus que sectorielles, les personnes interrogées reconnaissent très majoritairement l'engagement subjectif des chevaux dans le travail. Les locuteurs présentent leurs animaux comme des acteurs du travail capables d'initiatives et de négociations. Ce qui fait également consensus dans les représentations sociales des locuteurs c'est l'absence de différences dans le travail avec les chevaux selon que le professionnel soit un homme ou une femme (classe 16). Cela met au jour une nette évolution des représentations où la féminisation des activités équestres apparaissait alors comme un risque pour la pérennité de l'utilisation des chevaux (Tourre-Malen 2006 ; Digard, 2007). Bien qu'il existe une faible variabilité, la classe 13 qui est en lien avec la classe 5 indique un consensus autour de la figure idéale typique de l'homme de cheval. Elle est une valeur clé de chaque métier anthropoéquin. L'homme de cheval apparait comme le garant en acte de l'éthique professionnelle (Deneux – Le Barh, 2021b). Cette valeur est particulièrement prégnante quand il s'agit d'assurer de bonnes conditions de vie puis de gérer la retraite et de la fin de vie des équidés (classe 7 et 18), il s'agit d'être à la hauteur de l'engagement subjectif des chevaux dans le travail (Deneux – Le Barh, 2021c) dans une logique de don – contre don tel que théorisée par Porcher à partir des travaux de Mauss (Porcher, 2002).

4. Conclusion

L'analyse lexicométrique du corpus réalisée à partir du logiciel Iramuteq s'est centrée sur la classification hiérarchique descendante selon la méthode Reinert. Les mondes lexicaux stabilisés qui en résultent rendent compte des représentations sociales que les enquêtés des différents domaines professionnels ont quant à leurs relations de vie et de travail avec leurs chevaux. Le test d'association du chi2 de la variable domaine aux classes lexicales permet de mettre au jour le fait que la segmentation des « mondes du cheval » tient plutôt en la matérialité du travail avec les animaux. La motivation du professionnel à intégrer un métier anthropoéquin, la reconnaissance de l'agentivité des chevaux dans le travail et les engagements moraux que les humains ont envers leurs animaux sont communs à l'ensemble des groupes professionnels enquêtés.

Bibliographie

Deneux – Le Barh, V. (2021a). *La profession anthropoéquine : une identité marquée par une communauté socialisatrice de travail interspécifique* (thèse de sociologie). Université Paul Valéry, Montpellier III, France.

Deneux – Le Barh, V. (2021b). Travailler et vivre avec des chevaux. Confrontation de la moralité en acte des professionnels à l'économie morale animaliste. Communication présentée au colloque *La société morale* de l'AISLF, Tunis.

Deneux – Le Barh, V. (2021c). On n'achève plus les chevaux. *Géographie et cultures*, n° 115, p. 77-97.

Digard, J.P. (2007). *Une histoire du cheval. Art, techniques, société*. Arles : Actes Sud.

Porcher, J. & Estebanez, J. (eds.). (2019). *Animal Labor. A new perspective on human-animal relations*. Transcript Verlag.

Porcher, J. (2002). L'esprit du don : archaïsme ou modernité de l'élevage ? Éléments pour une réflexion sur la place des animaux d'élevage dans le lien social, *Revue du Mauss, n° 20*, p. 245-262.

Poudat, C. & Landragin, F. (2017). *Explorer un corpus textuel. Méthodes – pratiques – outils*, Louvain-la-Neuve, Belgique : De Boeck Supérieur.

Ratinaud, P. & 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,* 57-77.

Ratinaud P. (2009). *IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Texte et de Questionnaire* [programme informatique]. http://www.iramuteq.org

Reinert, M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 66, 5-39.

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.

Sapiro, G. (2007). La vocation artistique entre don et don de soi. *Actes de la recherche en sciences sociales*, 168, 4-11.

Tourre-Malen, C. (2006). *Femmes à cheval : la féminisation des sports et des loisirs équestres : une avancée ?* Paris : Belin.

Italian Institutional communication in pandemic period: a chronological analysis of Prime Minister speeches

Francesco Paolo Di Candia, Nicoletta Roberto,
Domenica Fiordistella Iezzi

University of Rome Tor Vergata - paolo.dicandia@gmail.com
nicoletta.roberto@gmail.com stella.iezzi@uniroma2.it

Abstract
Since the beginning of 2020, the whole world has been affected by the COVID-19 pandemic. In Italy, the first infections occurred in February 2020, immediately after the spread of the epidemic in China. In this frame, communication, especially that provided by official channels, has played a crucial role. Since the beginning of the pandemic, the YouTube Presidency of the Council of Ministers channel has been widely used in Italy. The paper aims is to study the mood of the speeches and compare it with the pandemic curve, in addition, the topics covered in the speeches are detected and observed according to the chronology. We analyzed 133 speeches of the Italian Prime Minister from February 2020 (the beginning of the spread of the virus in Italy) to September 2020 and to study how pandemic events influenced the language used in institutional discourse. In the time series of the contagion curve, we have identified three points of variation associated with four time periods. These periods are used to analyze grouped speeches using both an unsupervised approach based on cluster analysis and probabilistic topic analysis.

Keywords: chronological corpus, Italian Prime Minister speeches, LDA, sentiment analysis, topic analysis, predictive models.

1. Introduction

On January 9, the Istituto Superiore di Sanità (ISS) confirmed the first two cases of COVID-19 infection in Italy. To control the spread of the epidemic, subsequently classified as a pandemic, the government immediately activated a monitoring system to assess the associated health risk. From 24 February 2020, data relating to the spread of COVID-19 are collected, published and updated daily at 18:00 UTC by the Italian Civil Protection in a public repository which contains various parameters, including the number of infected, the number of

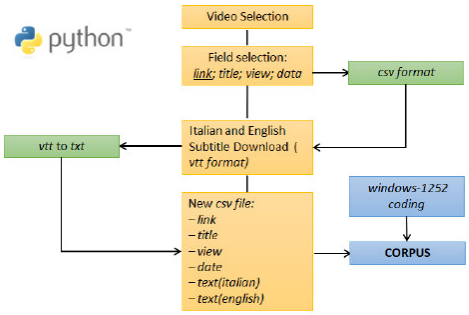
[1] Dipartimento Protezione Civile Italiana. Dati Covid-19 Italia. https://github.com/pcm-dpc/covid-19.

hospitalized in intensive care, the number of recovered, the number of tampons[1]. In this frame, communication has played a crucial role since the start of the pandemic. Information relating to COVID is constantly evolving. Therefore communication took place frequently through media with immediacy and wide diffusion as Social Media (SM), which were also used for communicating government decisions. In Italy official communication was often made of speeches given by the Italian Prime Minister, broadcast through institutional and traditional media (newspapers, TV, radio) and MS such as Facebook, Instagram, Twitter. From the beginning of the pandemic to today, Italy has seen a change of government. During the first pandemic period the President in office was Giuseppe Conte, and since February 2021 by President Mario Draghi. The main objectives of this work are:

- applying some of the main text mining techniques to the Prime Minister's speeches during the pandemic period (January 2020 - September 2021);
- studying the influence of the contagion curve on the language used in speeches, in particular, applying topic model technique to classify text over time;
- analysing the chronological trend of data and comparing them with the epidemic curve;
- experimenting with machine learning techniques to automatically classify text.

2. Data selection and preparation

All the institutional speeches given by the Italian Prime Minister and published on the YouTube institutional channel have been selected, from January 30, 2020, to September 30, 2021. The texts of speeches have been extracted through subtitles (vtt format[2]) either in Italian language or English language.



(a)

Corpus

Measure	*Value*
Row Number	133
Column Number	6
Languages	IT, UK
Footprint	3985684 bytes

Lexicometric

Measure	*Value*
Type	18418
Token	319892
Type Token Ratio	0.06
Hapax	8280
% Hapax	0.45

(b)

Figure 1. (a) Diagram to obtain corpus from YouTube Video (b) Corpus and lexicometric measurements

[2] Leo Zhou. vtt2txt.py convert youtube subtitles (vtt) to human readable text.

Texts in *vtt* were transformed in txt format to get corpus, as shown in Figure 1 (a). In order to obtain a structured data, the corpus was prepared by applying, at first, some simple normalization phases that consist of: elimination of apostrophes and punctuation, Windows-1252 decoding, conversion to lowercase letters, elimination of numerical digits, removal of "stopwords" via dictionary and searching stopwords through an empirical analysis of the frequency histograms of words in the text. Then on corpus were also applied lemmatization, in which each word is brought back to the basic form (in this work were used R Libraries : *UDPipe*[3] for the Italian text and tm for the English text), and or *stemming* in which each word is reduced from its inflected form to its root form called theme, by using R Library tm[4]. A set of lexicometric measurements on corpus in Italian is shown in Figure 1 (b). Type Token Ratio less than 0.2 and a percentage of hapax concerning the type of less than 50% (in this case, 45%) indicates that the text has a lexical richness and variety of words. These results allow corpus to be treated statistically.

3. Results

The daily trend of new positives, averaged over weekly values, is treated as a time series to identify some characteristics of the epidemic curve. In this curve, we detect four periods, as shown in Figure 2. The result of a cluster analysis is shown in 3; four clusters are depicted associated with the four pandemic periods.

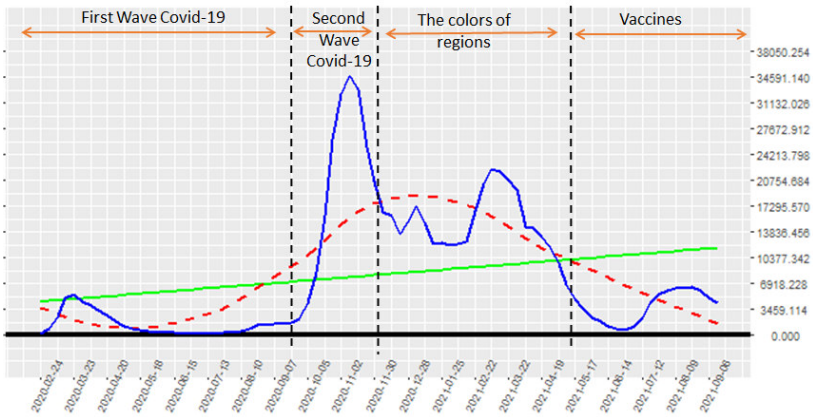


Figure 2. Time series of new positives averaged over weekly values (blue line), regression line (green) and LOESS curve (dashed red line)

[3] Wijffels, J. (2022). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.9. https://CRAN.R-project.org/package=udpipe

[4] Feinerer, I., Hornik, K. and Meyer D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5): 1-54. URL: https://www.jstatsoft.org/v25/i05/.

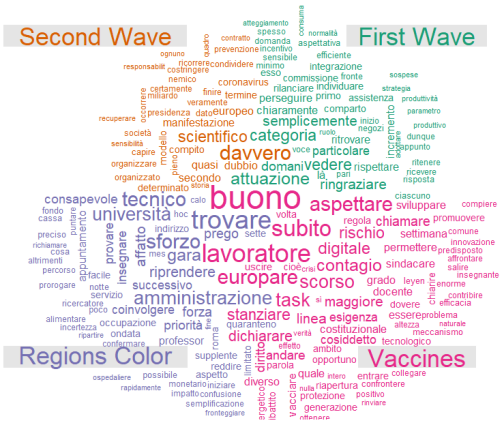


Figure 3. Wordcloud comparison

The comparison made using the R library[5], indicates that the language used has typical connotations in a certain period, which are less used in the others. It follows a list of terms for each period:

- “Assistenza (Assistance)”, “Perseguire (Pursue)”, “Attuazione (Implementation)” are frequently used in the *First wave* period;
- "Scientifico (Scientific)", "Coronavirus", "Prevenzione (Prevention)" are typical of the *Second Wave* period;
- “Sforzo (Effort)”, “Università (University)”, “Tecnico (Technician)” are evident in the *Colors of the Regions* period;
- "Lavoratore (Worker)", "Aspettare (to wait)" and "Contagi (Contagion)" are used frequently in the *Vaccines* period.

Furthermore, we extracted the sentiment of the texts using not only the binary classes (positive/negative) but also the emotional classes of the Plutchik wheel (Plutchik, 2001). In this work, a Lexicon Based approach is applied, based on R library Syuzhete for the English text and sentiment analysis stemming is applied to the corpus.

In Figure 4 emotions associate with each speech are showed using the Plutchik wheel; due to the lack of library in Italian language the emotions analysis is performed using texts in English language. The first period is characterized by *fear* and *trust*. Either between the first and, second period or at the beginning of the fourth period (when new positives were close to zero), *joy* and *surprise* were increased. In recently speeches, sentiment such as anticipation and joy are

[5] Fellows I. (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

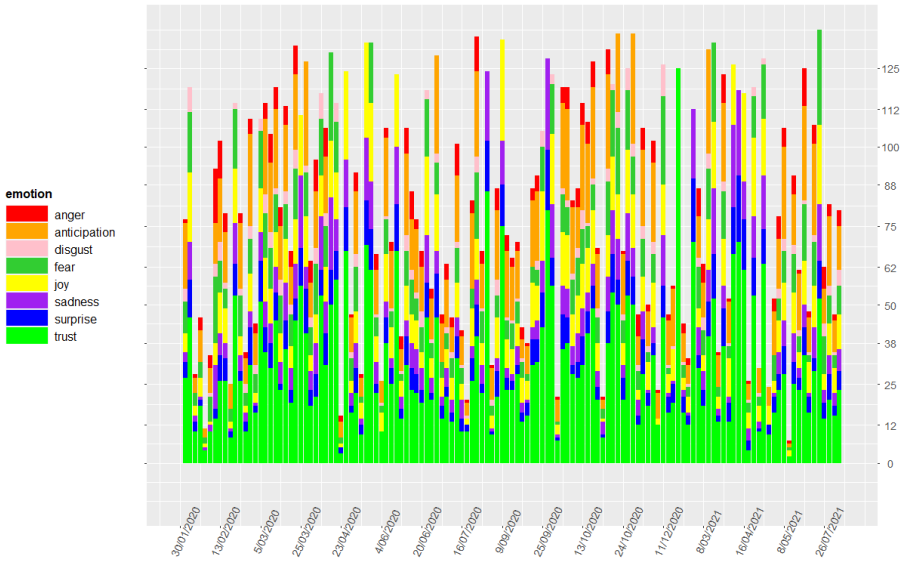


Figure 4. Time series of emotion associated to each speech

prevalent. Finally, to detect the topics in the corpus, we applied the Latent Dirichlet Allocation (LDA). LDA is a natural language analysis model that allows to understand the semantic meaning of the text by analyzing the similarity between the distribution of terms of the document with that of a specific topic. In this work by examining the semantic likelihood of the words present in topics with topical issues in the epidemic period. According to Figure 5, observing the trend over time of each speech belonging to the five topics interesting results are obtained:

- the speeches belonging to the topics "*Emergency management during the pandemic period*" and "*School, Work and Families during the pandemic period*" are very frequent until the end of 2020, which corresponds to the first period of the management of the pandemic in which measures to contain the epidemic and measures to support families and workers were issued very frequently;
- the speeches belonging to the topic of "*Vaccines*" are very frequent in the period which starts from the first dosing of the vaccines in Italy (February 2021) in line with government strategies;
- the speeches belonging to the topics "*Situation of Italy in the European and international pandemic context*" and "*Recovery plans from the pandemic period*" are present throughout the epidemic period.

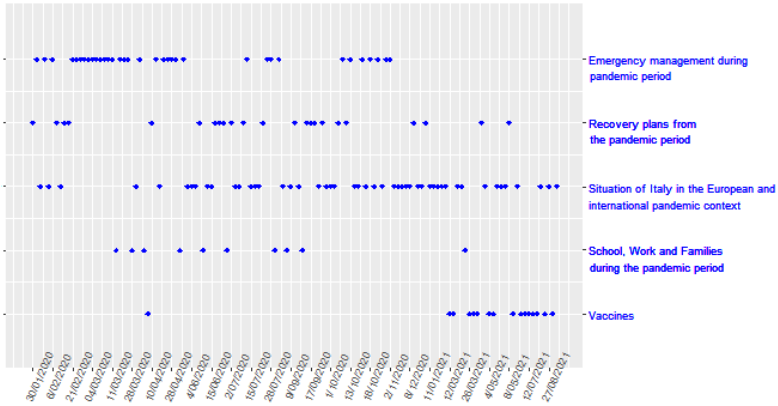


Figure 5. Temporal trend of the affiliation of the speeches to the five topics created

3.1. Use of topics for training predictive models

In order to evaluate an automatic classification of the texts belonging to a certain topic, it has been explored the use of some Machine Learning models. As a first trial, it was decided to instruct a single learning machine in order to be able to classify whether or not a discourse belongs to a single topic. The topic chosen to train the predictive model was the most numerous "*Situation of Italy in the European and international pandemic context*". In this training phase, the two programming languages R and Python were used. In particular:

- a document terms table has been obtained; on the rows are present the speeches of the Prime Minister while on the columns are present the words contained in the topic; the weights consist of the TFIDF index to enhance the most frequent graphic forms only in some papers; the features contained in multiple topics are present only once;
- the training was done using either neural network model (using Tensorflow in Python) or Support Vector Machine (using Sklearn in Python);
- the dataset was separated in a training set (80%) e test set (20%).

Table 1. Classifier confusion matrix

Type	*Accuracy*	*Precision*	*Recall*	*F1*
SVM	0.8148	0.8889	0.6667	0.7619
NN	0.8519	0.9	0.75	0.8182

From the examination of the values shown in Table 1, good results are obtained with both classifiers. Although the computational effort is greater in training the Neural Network on the other side with the Neural Network slightly more performing values are obtained either for accuracy or the F1 parameter.

4. Conclusions

This work highlights a general relation between text analyses and pandemic event. In particular, the following main results are highlighted:

- with the analysis of the clusters is possible to recognize four distinguishable groups associated with the four pandemics periods; frequently words found throughout the four periods are often typical keywords of the moment;
- encouraging results are obtained either from the probabilistic analysis of the topics or using the results of the probabilistic analysis to train a predictive model for automatic classification of speeches;
- the extraction of emotions from English text by using the classification of Plutchik's wheel produces interesting results in the analysis of the trend of emotions over time, particularly a great variability in relation to contagion curve for *trust*, *fear* and *anger* is detected.

The results (specially the machine learning approach used in topic classification) encourage to carry on further investigations to apply the methodologies developed in this work on different data-sets.

References

Plutchik R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4), 344-350.

Jockers M. L. (2015). Syuzhet: Extract sentiment and plot arcs from text. *Retrieved October,* 21, 2015.

Chauhan U. and Shah A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys* (CSUR), 54(7), 1-35.

Albalawi R., Yeap T. H. and Benyoucef M. (2020). Using topic modeling methods for short-text data: A comparative analysis. Frontiers in Artificial Intelligence, 3, 42.

Blei D. M., Ng A. Y. and Jordan M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Estimating Uncertainty in Narrative Economics Using Latent Semantic Scaling

Sami Diaf, Florian Schütze

Universität Hamburg – sami.diaf@uni-hamburg.de
florian.schuetze@uni-hamburg.de

Abstract
The study of economic narratives using text mining techniques increased recently with a dominance of unsupervised models built upon the bag-of-word assumption, to extract meaningful information for further econometric inferences. Yet word independence remains a constraint when the message relies on specific and intentionally word usage to explicit a particular message that often comes embedded with an ideological purpose. Monetary policy, as an important field of economics, carries uncertainty in its decision-making process as a forward guidance tool and an important aspect of its communication strategy. Noticeable is the construction of uncertainty indices using uncertainty-related word counts from different text sources. These are often embedded with an information bias that may bifurcate the intended aim of central banks. This work tries to quantify monetary policy uncertainty at the source and identifies its key drivers using *Latent Semantic Scaling*, a semi-supervised positional approach that renders similar capabilities as for word embedding, based on selected seed words given as priors to an augmented matrix factorization applied at the sentence level. Results show a receding uncertainty level at the Federal Reserve (1996-2020), due to a significant change in the monetary discourse after the 2008 financial crisis, where the use of uncertainty and risk-related words was reduced as part of a communication strategy aimed at sending neutral policy signals to the public amid unclear economic outlooks.

Keywords: document scaling, semi-supervised learning, uncertainty.

1. Introduction

The digitalization of text sources proved to be an invaluable source in linguistics to study the semantics. Economists embarked in the Narrative Economics wave (Shiller, 2017) and have been using various document sources to investigate further hypotheses and augment existing inferences with text-based covariates. Yet the dominant approach used so far belong to the topic model family, relying on word independence as for the Latent Dirichlet Allocation (Blei, 2003).
Several attempts tried to build specific measurements and gauge the content of corpora, as for sentiment analysis whose goal is to quantify the polarity of a

document based on available wordlists. Aside from the negative-positive sentiment duality, researchers tried to assess other sentiments or emotions as for uncertainty to have an in-depth insight into the corpora.

The *Economic Policy Uncertainty* index (Baker *et al*., 2016) remains a popular metric used by economists to assess uncertainty from a policy-oriented perspective, although built upon a simple word count methodology over journalistic corpora, prone to information biases. Other extensions are either specific purposes as for monetary policy (Husted *et al*., 2020) or country-specific. In monetary policy, uncertainty had been a feature of central bankers so to form a kind of monetary mystique (Goodfriend, 1986). Even if global transparency efforts at the end of 1980s brought more light to modern monetary policy practices (Gentzkow, 2019), practitioners still consider uncertainty as an integral part of monetary decision-making (Greenspan, 2004) and researchers tried to assess this direction, from a textual perspective, to shed more light on the inner functioning of monetary committees around the world.

This paper proposes a new approach to quantify monetary policy uncertainty related to specific words bearing the polarization of the discourse. As a semi-supervised model, *Latent Semantic Scaling* (Watanabe, 2020) helps identifying sentences related to a specific task by scaling them using a matrix-factorization approach that yields latent factors. Further aggregations help building an index based on these latent factors.

We argue that a dedicated factorization approach, at the sentence-level, proves to be powerful to identify implicit uncertainty better than a naïve word count or using a sentiment dictionary for a general assessment that might be hard to get. Moreover, this methodology tackles the shift in central banking communication practices after the financial crisis of 2008, where monetary policy committees around the world were not keen to use the "uncertainty" words, but prefer "risk" (Cieslak, 2021) probably to avoid any negative communication side effects to the economic agents.

The use of semantic scaling enables to identify and polarize context words related to uncertainty which inform us about the rhetoric style adopted by a given author, so to tackle two stylized facts found in central banking communication: ambiguity (Baerg, 2020) and consensus (Meade, 2005).

2. Scaling Models

Scaling techniques are mostly unsupervised methods used to rank documents on a given scale to facilitate comparison or clustering purposes. Depending on their assumptions, we can distinguish two families of scaling methods: non-probabilistic and probabilistic methods. Non-probabilistic approaches are usually related to sentiment analysis, as users employ dictionaries or wordlists to given documents scores based on pre-established weightings, while probabilistic

approaches are mostly borrowed from statistical learning and infer document positions based on an optimization scheme, used to retrieve one or more hidden dimensions or scales. *Wordscores* (Laver *et al*., 2003) uses a user-defined wordlists that meet the needed dimension.

Wordfish (Slapin *et al*., 2008) uses a Poisson model on word counts to extract a scale of document positions, assumed to follow a normal distribution, built upon word/document contributions. This scale could be used as a proxy of an ideological dimension and further extensions are possible to make aggregation at the metadata level (Lauderdale *et al*., 2016; Diaf *et al*., 2021).

Latent Semantic Scaling (Watanabe, 2020) (LSS) offers a semi-supervised approach for scaling, where users could focus on a specific dimension related to the occurrence of selected words, whose surrounding words (context) have an informative power similar to distributional representations (Mikolov *et al*., 2013), known as word embedding, which aim to vectorize words according to their context.

These surrounding words are corpus specific while the selected words (or seed words) are not. For example, the wording of central bank communication could be highly specific when it comes to explicit uncertainty about monetary policy; therefore, it is not sufficient to just use a generic uncertainty dictionary. Therefore, the semi-supervised LSS approach combines exogenous seed words with terms that narrowly identifies textual distance from seed words, thus, taking advantage of both probabilistic and non-probabilistic approaches.

3. Application and Results

Governors' speeches at the Federal Reserve constitute a reliable source to assess external monetary policy communication, where central bankers speak about topics pertaining to the US economic and monetary status. FOMC transcripts are released with a five-year delay and inform about the debated themes during FOMC meetings, although consensus and ambiguity emerge as stylized facts from this type of documentation. The corpus of governors' speeches spans from January 1996 to December 2020, with 1,488 documents given on several occasions at different locations by 26 different speakers. The mean word count within a document is 3,235 words with a standard deviation for all documents of 1536. Furthermore, in each year on average 60 speeches were hold with a standard deviation of 17. Pre-processing steps as for the removal of numbers, English stopwords, punctuation, symbols, URLs and separators were conducted to build a document-feature matrix, as a ready-to-use input for scaling estimation.

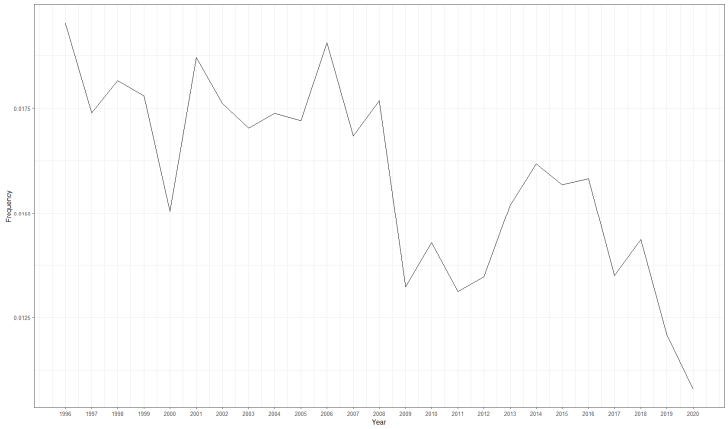


Figure 1. Frequency of exogenous uncertainty words (from Loughran-McDonald dictionary) in the overall corpus.

Loughran-McDonald sentiment dictionary (Loughran and MacDonald, 2011) gathers 297 different words related to uncertainty. In Figure 1, the frequency of Loughran-McDonald uncertainty words over time can be seen with an evidence of a downward trend regarding the frequencies in governor speeches indicating avoidance of uncertainty terms. This uncertainty dictionary will be used as seed words for LSS estimation to get significant terms in a window of three context words, at a significance level of 5%, depending on their proximity to a lexicon of positive and negative words (Hu and Liu, 2004). It results 1,245 words used as polarizing words for the LSS scores that will be estimated using 300 latent factors. These 300 latent factors are the default number of hyperparameters suggested by Watanabe (2020).

The smoothed LSS index using LOWESS, as demonstrated in Figure 2, has an overall downward slope, except during the financial crisis (2008-2010) meaning the word usage at the Federal Reserve was meant to be reassuring, as a strategy to dissipate worries about economic/monetary risks. This fact demonstrates ambiguity as a feature of modern central banking communication, avoiding any false signal sent to economic agents that could trigger a potential negative impact, even in times of economic downturn.

Previous crises that hit the United States, as for the Asian crisis (1997) and the internet bubble (2001) had been dealt by Federal Reserve Governors with a great uncertainty, if compared with subprime crisis (2007-2008) and the debate following the quantitative easing in 2010. One potential explanation of this shift could be the digitalization and the accelerated transparency efforts embedded with an informative, but neutral, communication strategy widely adopted by central banks around the world. Nevertheless, the LSS index remains a bias-free measurement of how uncertain the monetary rhetoric is, if compared to other

indices from newspapers and magazines that are often skewed with partisanship and ideological stance.

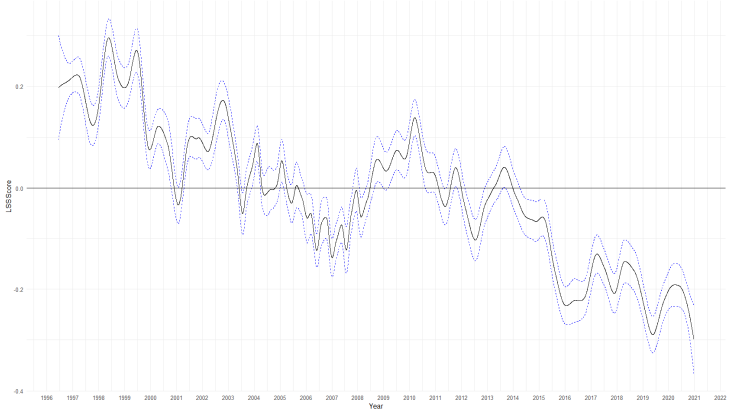


Figure 2. LSS-based uncertainty index (smoothed with LOWESS) and its 95% confidence interval.

When comparing the frequency of Loughran-McDonald uncertainty words (Figure 1) and LSS-based uncertainty (Figure 2) the second one brings more details, not only regarding the extent of uncertainty but also about the sentiment related to this uncertainty. A higher score could result from the usage of more uncertainty context words and/or due to the usage of more context words with a positive sentiment. LSS estimation provides polarity scores for context words related to uncertainty seeds, which have a small weight in absolute value. Figure 3 shows that words contributing to elevated Uncertainty scores are mostly used for negative sentiments which is in line with theory proposing uncertainty per se is a negative economic driving force.

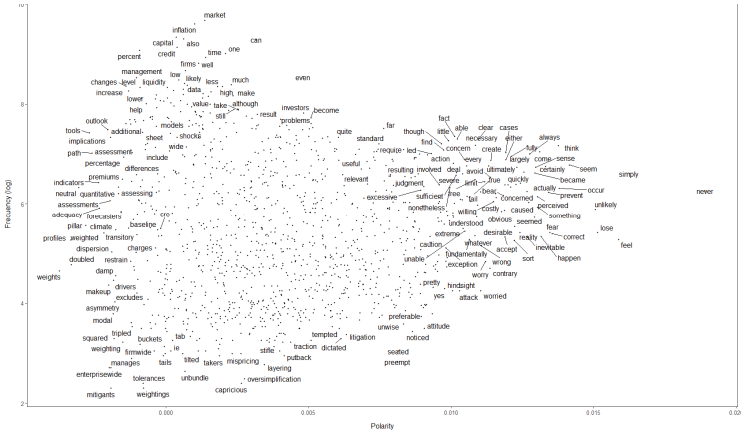


Figure 3. Most polarized words estimated with LSS

4. Conclusion

Aside from traditional sentiment analysis and usual wordlists employed to construct text mining indices, the use of semi-supervised models offers wide possibilities to overcome the constraining bag-of-word hypothesis and allow a sound, lower-level approach whose power lies on identifying context words related to the study. Latent Semantic Scaling relies on user-provided seed words and a latent semantic analysis whose performance was found similar to distributional representations that employ neural networks.

LSS yields a smoothed, downward uncertainty-index, in-line with the shift in the rhetoric used by Federal Reserve Governors in the United States (1996-2020) which witnessed a lower-level use of uncertainty-related words after 2008 the financial crisis. Context words, which bear the importance of the uncertainty index, have small coefficients in absolute value, with positive polarization is likely to concern negative words. The resulting polarization words could be used as a wordlist for a comparable study in monetary policy uncertainty to establish if the communication of different central banks follow the same direction as the Federal Reserve.

References

Baker S.R. and Bloom N. and Davis S.J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Baerg N. (2020). Crafting Consensus: Why Central Bankers Change their Speech and how Speech Changes the Economy. *Oxford University Press*.

Blei D.M. and Ng A.Y and Jordan M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research* (3): 993–1022.

Cieslak A. and Hansen S. and McMahon M. and Xiao S. (2021). Monetary policymakers' uncertainty. ECB Conference on Monetary Policy: bridging science and practice.

Diaf S. and Döpke J. and Fritsche U. and Rockenbach I. (2022): Sharks and minnows in a shoal of words: Measuring latent ideological positions based on text mining techniques, *European Journal of Political Economy*.

Gentzkow M. and Kelly B. and Taddy M. (2019). Text as Data. *Journal of Economic Literature*, 57 (3): 535-74.

Goodfriend M. (1986). Monetary mystique: Secrecy and central banking. *Journal of Monetary Economics*, 17(1): 63-92,

Greenspan A. (2004). Risk and Uncertainty in Monetary Policy. *American Economic Review,* 94 (2): 33-40.

Hu M. and Liu B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). Association for Computing Machinery, New York, NY, USA, 168–177.

Husted L. and Rogers J. and Sun B. (2020): Monetary policy uncertainty, *Journal of Monetary Economics*, 115, 20-36,

Lauderdale B. and Herzog A. (2016). Measuring Political Positions from Legislative

Speech. *Political Analysis*, 24(3): 374-394.
Laver M. and Benoit K. and Garry J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2), 311-331.
Loughran T. and McDonald B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66: 35-65.
Meade E.E (2005). The FOMC: Preferences, Voting, and Consensus. Federal Reserve Bank of St. Louis Review, 87(2): 93-101
Mikolov T. and Sutskever I. and Chen K. and Corrado G.S and Dean J. (2013) Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26:
Shiller R.J. (2017). Narrative economics. Working Paper 23075, National Bureau of Economic
Research.
Slapin J. B. and Proksch S.O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52(3): 705-722.
Watanabe K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, 15(2): 81-102.

Cross-lingual Macroeconomic Narratives: Different Views Of The Same Problem?

Sami Diaf[1], Rachid Toumache[2]

[1]Universität Hamburg – sami.diaf@uni-hamburg.de
[2]Ecole Nationale Supérieure de Statistique et d'Économie Appliquée (ENSSEA) – rtoumache@gmail.com

Abstract
Narrative economics triggered a great interest of economists on the use of text as data to test further hypotheses and augment existing econometric inferences. While most of the text-based applications were designed for monolingual corpora, it often happens that documents collected in different languages offer a complete information source but not fully exploited using usual methods. This work tries to understand partisanship in macroeconomic assessment using narratives of the International Monetary Fund and a country member (Algeria) to identify potential differences contained in their published reports by applying a cross-lingual contextualized topic model with ZeroShot learning which ensures topics to be language-transferable and projected to documents in different languages, so to endow a robust, semantic-based transfer mechanism for a joint document analysis. Results show the transfer learning process yields semantically coherent, dynamic topics that meet the macroeconomic diagnostic advocated by the IMF, while the narrative structure at the Bank of Algeria was found to be less dynamic, describing government efforts to consolidate the macroeconomic resilience and reduce vulnerabilities. Particularly, recent IMF's advocacy seems to target fiscal policy, whose implementation does not meet the central bank's scope.

Keywords: topic models, transfer learning, macroeconomics.

1. Introduction

Narrative economics (Shiller, 2017) established itself as a growing subfield of economics, benefiting from the continuous developments of machine learning techniques that enable an in-depth analysis of textual data. Yet the dominant technology used by economic researchers remain topic models, which try to learn hidden representations in a form of groups of words, known as topics. As for the Latent Dirichlet Allocation (Blei *et al*, 2003), considered as the go-to solution, topics are usually not semantically structured and concerns about topic quality remain subjected to the experience of researchers who often proceed with post-hoc inferences or using priors as semi-supervised augmented methods (Eshima

et al, 2020). Most important, the proposed algorithms fit monolingual corpora and their application to multilingual documents remains problematic, as they require a transfer mechanism that translates the results from one language to another (Bianchi *et al*., 2021), often done manually or via machine translation tools (Lucas *et al*., 2017).

Text narrative attempts at the IMF targeted country-based uncertainty indices (Ahir *et al*., 2022), growth narratives (Cherif *et al*., 2020) and IMF's priorities (Anderson *et al*., 2021), however, little was done concerning country surveillance, although it remains an important pillar of the Fund's scope, aiming at providing consultancy on risks and vulnerabilities that may threaten members' economic stability. In many cases, country members do not share the same views of the IMF staff on the economic assessment and proceed by issuing reports or communiqués highlighting differences with the staff's assessment. Such debates identifies key differences in macroeconomic narratives, however, many countries use their national languages instead of the standard English language adopted by the IMF staffs, although recent IMF publications were given in many languages so to reach a wider audience.

This work proposes to study a macroeconomic debate through textual narratives, by adopting a transfer learning approach built upon pre-trained multilingual models, capable of analyzing documents written in different languages and making a fair assessment about the debated topics and common interests. As an illustrative application, we use IMF staff reports (*Article IV*) issued during the period 2002-2018 for the case of Algeria, an oil-exporting IMF-country member, along the annual reports issued by Algerian Central Bank (*Bank of Algeria*) in French. Cross-lingual modeling (Bianchi *et al*., 2021) will provide a tool to highlight inter-language similarities and differences that cannot be captured by traditional text analytics' tools by learning topics in rich-resource language (English) documents and then predicting topic scores for another language (French) to have a common framework for comparative purposes.

2. Multilingual Topic Models

Topic models constitutes a family of mostly unsupervised models used to study monolingual corpora, whose aim is to uncover groups of words with a semantic meaning, known as topics. Besides the Latent Dirichlet Allocation (Blei *et al*., 2003), several developments were proposed to tailor the study to the corpora specifications, but fail to get semantic and coherent topics, as they rely on the bag-of-word hypothesis.

The resurgence of distributional representations in natural language processing (Mikolov *et al*., 2013), also known as *word embedding*, proved to be a powerful solution to many natural language processing technologies as it maps words into dense vectors that encompass the semantic and syntactic information from their

occurrences in the corpus. Further extensions consider mapping documents into vectors (Le *et al*., 2014) and labelling context (Devlin *et al*., 2019) that could be applied to a variety of tasks.

While the use of standard topic models for multilingual corpora remains constrained as translating becomes problematic for polysemous words, pre-trained multilingual models could be used to link topics of different languages (Bianchi *et al*., 2021). Such transfer learning mechanisms were found to carry a high topic coherence and stable results in comparison with other proposed technologies as it relies on an augmented input mixing bag-of-word and a multilingual contextualized embedding. Such neural topic modeling architectures are able to hold contextual information to yield a coherent output in terms of topics, in comparison with traditional neutral architectures that only use bag-of-word representations. Moreover, ZeroShot CTM infers topics on unseen documents in other languages than the training one, in addition to the ability to handle unseen words in the test set that are not present in the training set (Bianchi *et al*., 2021) by calling a multilingual pre-trained model.

3. Application and Results

The corpus consists of the IMF's Article IV reports for Algeria as well as the Bank of Algeria's annual reports (BA) during the period 2004-2018. Scanned documents were digitalized and cleaned to serve as an input to our study.

Bank of Algeria's annual reports are a reliable source of information about the Algerian economy as it compiles, in addition of monetary and financial data, other information about unemployment, budget and real sector statistics in a chapter-structured reports augmented with tables in the appendix. The BA reports are released in French with a one-year delay, compared to the timely Article IV reports, except for the years 2012 and 2015.

We apply the Zero-Shot Contextualized Topic Model (Bianchi *et al*., 2021) on the joint corpus of the IMF and BA reports, where the corpus of the IMF will be used to estimate the model in English language (training set), and then serves as a basis to predict the topics' prevalence on the French language for the BA documents (test set). The number of topics to be learned, a hyperparameter, was chosen to be ten to encompass the information contained in the documents and have a coherent, non-overlapping topic structure able to capture narrative similarities.

Table 1 shows the learned topics on the IMF corpus with a dominance of abbreviations and technical terms related to hydrocarbons, budget and monetary. Topics 2, 6 and 9 deal with international environment, while the remaining concern liberalization, non-hydrocarbon (nrc) growth and fiscal policy. Topics 5 and 7 share similar top 10 words although used in different contexts, potentially in different reports, linked to the banking sector (Topic 5) and to staff projections

(Topic 7). The learned topic structure is coherent and meets the different issues IMF were raising during its staff missions as for diversifying the economy and foster the non-hydrocarbon sector, reducing unemployment and gender disparities. Noticeable is the presence of specific terms related to projections as for prel (preliminary), baseline and deviations.

Table 1. Top 10 words of topics learned by Zero-Short CTM on the IMF documents.

Topics	Top 10 words
1	true, however, privatization, restructuring, deviations, institutional, tariff, standard, treasury, part
2	Percent, fiscal, debt, gdp, international, market, financing, monetary, baseline, nonhydrocarbon
3	adjusting, nhc, accelerates, liberalize, starting, median, window, adjusted, prel, appreciate
4	figure, article, expenditure, iv, social, revenues, pressure, salaries, unemployment, measures
5	adjusting, adjusted, nhc, accelerates, median, entry, male, test, basel, ownership
6	percent, fiscal, gdp, debt, international, rate, growth, oil, monetary, algeria
7	adjusting, adjusted, nhc, accelerates, median, male, pressure, falls, expectations, basel
8	rule, hydrocarbon, financial, consumption, markets, revenue, spending, women, development, algeria
9	percent, debt, fiscal, gdp, international, monetary, rate, nonhydrocarbon, market, financing
10	public, authorities, projections, hydrocarbon, us, external, debt, banks, banking, system

The learned ZeroShot CTM on the IMF's document (training set) was used to predict the topic scores on the BA documents (test set) and then compare the most prevailing topic in each institution's report, as demonstrated in Table 2. IMF narratives are clearly dynamic and cover different macroeconomic areas compared to the BA reports, which have an overall descriptive structure, in terms of reporting, dominated by two topics (5 and 7). IMF staff reports show a special biennial interest, which could be linked to the staff appointments. Except for the years 2008 and 2009 where a somewhat alignment in the narratives was found due to the prevailing financial crisis, there is a clear departure in terms of topic preference at both institutions. Moreover, IMF staff reports for the years 2016-2018 seem to target fiscal policy and internal debt (Topics 2 and 9) which does not fully meet the scope of the BA, although detailed budget statistics were given on the BA's annual reports.

Table 2. Most dominant topics per institution per year.

Dominant Topic BA	Year	Dominant Topic IMF
5	2004	10
5	2005	10
7	2006	3
7	2007	3
5	2008	5
7	2009	7
5	2010	7
7	2011	4
7	2012	NA
7	2013	8
7	2014	8
5	2015	NA
5	2016	9
5	2017	9
7	2018	2

4. Conclusion

Most of machine learning algorithms for text data were designed for monolingual corpora, but recent improvements in natural language processing endow practitioners the possibility to study multilingual corpora and test cross-lingual hypotheses. In macroeconomics, overall assessment has been approached via different views and identifying partisanship or alignment remained difficult to proof, especially with multilingual documents. This study builds a cross-lingual framework to compare macroeconomic narratives at the IMF and at a country member (Algeria) to assess a potential partisanship or differences in the views that explain the debate about the Article IV assessments. Findings show a dynamic topic structure contained in the IMF staff reports targeting non-hydrocarbon sector, fiscal policy and the real sector, while Bank of Algeria's annual reports seem to have a descriptive narrative whose structure is not aligned to the views of the IMF, except during the 2008 financial crisis, where a consensus in the views was prevailing.

References

Ahir H., Bloom N. and Furceri D. (2022). The World Uncertainty Index. NBER Working Papers, 29763.

Anderson G., Galang P., Gamba A., Medina L. and Zheng T. (2021). How Have IMF Priorities Evolved? A Text Mining Approach. IMF Working Papers, 21/159.

Bianchi F., Terragni S., Hovy D., Nozza D. and Fersini E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: 1676–1683

Blei D.M., Ng A.Y and Jordan M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research* (3): 993–1022.

Cherif R., Engher M. and Hasanov F. (2020). Crouching Beliefs, Hidden Biases: The Rise and Fall of Growth Narratives. IMF Working Papers, 20/228.

Devlin J., Chang M., Lee K. and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (1): 4171–4186.

Eshima S., Imai K. and Sasaki T. (2020). Keyword Assisted Topic Models. Working Paper, arXiv:2004.05964.

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A. and Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. Political Analysis, 23(2), 254-277.

Mikolov T., Sutskever I., Chen K., Corrado G.S and Dean J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26:

Le Q.V. and Mikolov T. (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on International Conference on Machine Learning, (32): II-1188–II-1196

Shiller R.J. (2017). Narrative economics. Working Paper 23075, National Bureau of Economic Research.

Présence des noms d'agents au féminin dans un corpus de presse. Étude comparée Québec-Belgique.

Anne Dister[1], Hubert Naets[2], Patrick Watrin[2]

[1]Université Saint-Louis – Bruxelles – anne.dister@usaintlouis.be
[2]Université Catholique de Louvain – hubert.naets@uclouvain.be
patrick.watrin@uclouvain.be

Abstract
In this article, we are interested in the implementation of feminine names of professions, titles, ranks and functions in the press. We analyse two French-language daily newspapers, one Belgian (*Le Soir*) and the other Quebecois (*Le Devoir*) at a 20-year interval (2001 and 2021). The results we obtain show that female names are much more numerous in 2021, but also indicate a difference between Quebec and Belgium.

Keywords: feminization, corpus, diachrony, Belgium, Quebec, Unitex

Résumé
Dans cet article, nous nous intéressons à l'implantation des noms de professions, titres, grades et fonctions au féminin dans la presse. Pour ce faire, nous analysons deux quotidiens francophones, l'un belge (*Le Soir*) et l'autre québécois (*Le Devoir*) à 20 années d'intervalle (2001 et 2021). Les résultats que nous obtenons montrent que les noms au féminin sont beaucoup plus nombreux en 2021, mais indiquent aussi une différence entre le Québec et la Belgique.

Mots clés: féminisation, corpus, diachronie, Belgique, Québec, Unitex.

1. Introduction

La féminisation des noms de professions, titres, grades et fonctions est loin d'être un sujet de d'études récent, que ce soit du point de vue social ou linguistique. C'est dans les années 1970 que le Québec légifère sur la question, suivi en cela par les autres pays de la Francophonie Nord (pour un aperçu historique, voir Dister et Moreau 2009). Le principe est simple, et guidé par le souci de visibiliser les femmes dans la sphère professionnelle, conformément à la morphologie du français et au fonctionnement de la langue : on préconise d'utiliser un nom au féminin pour désigner une ou des femmes. On dira ainsi *Madame l'ambassadrice* et non **Madame l'ambassadeur*, *une avocate* et non **un avocat*, *une juge assermentée* et non **un juge assermenté*. Les décisions politiques ont la plupart

du temps été suivies de la publication de guides et encouragées par des personnalités publiques qui ont revendiqué un titre au féminin. La presse est, pour la question qui nous occupe, un bon observatoire des pratiques, comme l'a montré Fujimura (2005) qui a étudié l'évolution des noms d'agent au féminin dans deux journaux français, avant et après la prise de décision politique de 1998 par le gouvernement de Lionel Jospin. Dans cet article, c'est sur la presse belge francophone et québécoise que notre attention s'est portée, dans un perspective également diachronique.

2. Méthodologie

2.1. Collecte des données

Nous avons constitué notre corpus à l'aide du service *Europresse* (http://www.europresse.com/fr/), une base de données payante d'informations en ligne disponible sur internet. Nous avons récupéré manuellement tous les articles proposés par *Europresse*[1] pour le journal montréalais *Le Devoir* et pour le journal belge francophone *Le Soir* pour les années 2001 et 2021. Les deux quotidiens ont un lectorat et un positionnement assez comparables. Nous avons choisi, d'une part, l'année 2021, car il s'agit de l'année complète la plus récente et, d'autre part, l'année 2001, car il nous a semblé qu'un écart de vingt ans devrait permettre d'observer une évolution significative des usages.

Tableau 1. Détail des corpus collectés

	2001		**2021**	
	#Articles	**#Tokens**	**#Articles**	**#Tokens**
Le Devoir	24 939	13 617 976	14 549	10 152 902
Le Soir	48 800	21 394 200	26 949	11 328 943

2.2. Constitution des ressources

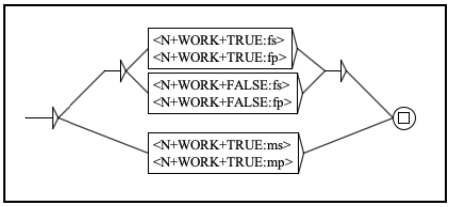
Afin d'identifier et d'extraire les phrases contenant un nom de métier repris dans Moreau et Dister (2014), nous avons utilisé Unitex (Paumier, 2021), une suite logicielle libre d'analyse de corpus. Cet outil repose sur la création de ressources linguistiques :

- des dictionnaires (Courtois et Silberztein, 1990) qui associent des informations morphosyntaxiques à chaque mot du texte ;
- des grammaires locales (Gross, 1997) qui formalisent des structures lexico-syntaxiques intégrant les mots et leurs informations morphosyntaxiques.

[1] Nous sommes tributaires des données rendues disponibles par Europresse, qui ne propose pas tous les articles publiés.

Outre les dictionnaires du français intégrés à l'outil, nous avons développé un dictionnaire de noms de métiers qui, à chaque entrée non épicène[2], associe des informations flexionnelles (de manière à permettre la distinction masculin/féminin) et un ensemble d'informations sémantiques. Ces dernières ont pour objectif d'identifier les noms de métier au sein de nos grammaires locales (étiquette *+WORK*) mais également de distinguer les formes recommandées (étiquette *+TRUE*) des formes construites artificiellement (étiquette *+FALSE*). En effet, pour permettre l'étude de la créativité lexicale, il nous fallait être à même d'observer l'ensemble des possibles, mais également de les distinguer lors de nos analyses. Notons, finalement, que nous avons séparé (dans deux dictionnaires distincts, filtré et non filtré) les formes féminines sémantiquement non ambigües des formes potentiellement ambigües (c'est-à-dire ne désignant pas spécifiquement un nom de métier, comme *couvreuse* ou *moissonneuse* qui désignent soit le métier soit l'outil).
Nos grammaires locales sont quant à elles rudimentaires dans la mesure où leur seule fonction est l'identification des formes reprises dans le dictionnaire des noms de métiers (*cf.* Figure 1).

Figure 1. Grammaire locale principale



2.3. Extraction des candidats

L'application des dictionnaires et de notre grammaire locale sur nos corpus[3] nous a permis d'extraire les phrases contenant un nom de métier. Certaines formes pouvant également être utilisées en tant qu'adjectif (culture *religieuse*, par exemple), nous avons utilisé un étiqueteur morphosyntaxique[4] afin d'identifier les formes nominales et d'ainsi faciliter l'analyse des résultats.
L'ensemble des phrases extraites a été stocké pour chaque corpus au sein d'un fichier CSV (*cf.* Table 2). Chaque ligne est composée de quatre colonnes spécifiant le lemme, la forme, la nature nominale de la forme ("1" si la forme

[2] Notre objectif étant l'observation de la féminisation des noms de métier, il nous a semblé difficile d'intégrer les formes qui ne marquent pas graphiquement le féminin. Ces formes ont donc été manuellement supprimées de la liste.
[3] Pour permettre l'intégration d'Unitex dans notre outil d'extraction, nous avons utilisé la librairie python-unitex (https://github.com/patwat/python-unitex).
[4] https://spacy.io/

est un nom et “0” dans le cas contraire) et la phrase d’origine. Si une phrase contient plus d’une forme de nom de métier, plusieurs lignes sont créées.

Tableau 2. Exemple d’extraction pour Le Devoir 2021

Lemme	Forme	Nom	Phrase
auteur	*autrice*	1	Après sa drôle et intime Année exemplaire, l'**autrice** marseillaise lance une structure d'édition plus favorable aux dessinateurs.
auteur	*auteures*	1	Diane Gistal mentionne des **auteures** afro-québécoises qui gagneraient à être connues : Stéphane Martelly ou Chloé Savoie-Bernard, par exemple.

3. Analyse des résultats

3.1. Analyse quantitative

Pour éviter tout biais a priori, nous nous sommes concentrés ici sur les formes féminines filtrées (sémantiquement non ambigües, *cf.* Section 2.2) et étiquetées comme nom par le tagueur (*cf.* Section 2.3). Les comptes sont rapportés dans la table 3[5]. Sur base de ces comptes, nous avons calculé différentes valeurs de chi-carré afin d’évaluer la tendance à la féminisation en fonction des journaux et des années.

Tableau 3. Comptes pour les corpus Le Devoir et Le Soir

	2001			2021		
	TRUE	FALSE	OTHERS	TRUE	FALSE	OTHERS
Le Devoir	3061	20	13614895	5661	357	10146884
Le Soir	3713	4	21390483	3325	179	11325439

Pour 2001, le chi-carré ($\chi^2(2)=132,8$) montre un lien fort entre la féminisation et le journal. En effet, si l’on regarde les résidus standardisés, on constate un emploi nettement plus important de noms de métiers féminisés au sein du journal Le Devoir (Z=10,6 pour TRUE ; Z=4,5 pour FALSE). Cette différence s’accentue plus encore en 2021 ($\chi^2(2)=973,5$; Z=29,9 pour TRUE ; Z=8,9 pour FALSE).
Par ailleurs, si l’on oublie les différences entre journaux (en fusionnant les comptes pour Le Soir et Le Devoir), on constate un lien très marqué entre l’année et la tendance à la féminisation ($\chi^2(2)=3227,1$). À nouveau, les résidus

[5] Rappelons que TRUE et FALSE désignent respectivement les formes recommandées par Moreau et Dister (2014) et les formes construites artificiellement. OTHERS désigne le reste des mots.

standardisés permettent d'apprécier un emploi de la féminisation nettement plus marqué en 2021 (Z=49,4 pour TRUE ; Z=28,1 pour FALSE).

3.2. La catégorie des FALSE

Nous avons appelé "FALSE" les formes féminines non préconisées par le guide de 2014 qui nous sert de base. Nous manquons de place pour une analyse détaillée, mais l'essentiel de ces formes sont des mots en -eure (*acupuncteure, sculpteure, contrôleure, vérificateure*…) et en 2021 la majorité des items est représentée par le féminin *autrice*, forme qui était tombée en désuétude et n'était pas recommandée par les guides de féminisation. Ce terme, parce qu'il est un féminin ostentatoire, marqué aussi à l'oral, revient en force dans l'usage.

3.3. Le cas du lemme chercheur

Nous nous penchons dans cette section sur les formes du lemme chercheur au masculin et au féminin, uniquement au singulier (Table 3). En effet, les emplois du masculin pluriel sont très majoritairement utilisés pour neutraliser l'opposition de sexe. Dans un corpus de près de 7 heures de conversation, Branca et Dister (à paraitre, 2022) relèvent que 98, 4% des emplois du masculin pluriel fonctionnent de cette manière. Les emplois au féminin pluriel sont quant à eux toujours exclusifs : ils ne permettent de désigner que des référents de sexe féminin, ce qui est le cas dans nos données.

Tableau 4. Occurrences des formes du lemme chercheur pour les corpus Le Devoir et Le Soir de 2001 et 2021

	Le Devoir				*Le Soir*			
	2001		2021		2001		2021	
	#	%	#	%	#	%	#	%
chercheur								
homme	340	69,7	376	51,6	291	78,9	224	55,6
femme (déterminant masculin)	3	0,6						
femme (déterminant féminin)							1	0,2
femme (apposition)	2	0,4			7	1,9		
générique	69	14,1	28	3,8	34	9,2	19	4,7
chercheuse	61	12,5	322	44,2	37	10	159	39,5
chercheure	13	2,7						
chercheur et chercheuse			3	0,4				

Qu'il s'agisse du *Devoir* ou du *Soir*, on constate que, entre 2001 et 2021, la proportion de *chercheur* pour désigner un homme diminue, tandis que celle de chercheuse augmente, ce qui semble correspondre au fait que le nombre de femmes chercheuses augmente et/ou que leur visibilité est plus importante. Cela va dans le sens de la tendance générale observée sur toutes les formes du corpus.

4. Conclusions

Notre étude n'avait pas pour ambition de comparer l'emploi du masculin et du féminin pour désigner des femmes, mais de voir l'implantation des termes au féminin dans un corpus de presse. De 2001 à 2021, on constate une évolution significative de l'utilisation des formes au féminin, tant en Belgique qu'au Québec. Les mots au féminin sont beaucoup plus nombreux dans les textes. Sans doute est-ce pour partie dû à l'évolution de la société en général, et des positions plus nombreuses qu'y occupent les femmes. Précurseur en la matière, le Québec faisait déjà plus de place aux femmes, à travers des mots au féminin, dans les données de 2001. Cette tendance générale s'observe dans l'analyse plus fine des emplois de *cherc*heur et *chercheuse*.

Bibliographie

Branca S. et Dister A. (à paraitre, 2022). Le genre grammatical des noms désignant des référents humains. Ce que nous montre l'analyse de données orales non planifiées, *Description de l'oral et méthodes d'analyse en linguistique.*

Courtois A. et Silberztein M. (1990). Dictionnaires électroniques du français. *Langue française, 87* : 3-4.

Dister A. et Moreau M.-L. (2021). Madame l'Administrateur, c'est presque fini… La dénomination des candidates lors des élections : étude diachronique. In Fagard B. et Le Tallec G. Éditeurs, *Entre masculin et féminin… Approche contrastive : français et langues romanes.* Paris, Presses de la Sorbonne Nouvelle : 31-54.

Dister Anne and Moreau Marie-Louise (2009). *Féminiser ? Vraiment pas sorcier ! La féminisation des noms de métiers, fonctions, grades et titres*, De Boeck Duculot, Bruxelles (coll. « Entre guillemets »).

Fujimura I. (2005). Politique de la langue : la féminisation des noms de métiers et des titres dans la presse française (1988-2001), *Mots. Les langages du politique*, 78 : 37-52.

Gross M. (1997). The Construction of Local Grammars. In Roche E. et Schabès Y. Éditeurs, *Finite-State Language Processing*. Cambridge, MIT Press : 329-354.

Honnibal M. and Johnson M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*. Lisbon : 1373-1378.

Moreau M. and Dister A. (2014). *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre* (3e édition). Fédération Wallonie-Bruxelles.

Paumier S. (2021), Unitex 3.3. Manuel d'utilisation. URL :
https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-fr.pdf

Les temps de la crise sanitaire au prisme d'une série chronologique: une étude phonético-textométrique

Sascha Diwersy[1], Ivana Didirková[2], Christelle Dodane[1], Fabrice Hirsch[1], Giancarlo Luxardo[1]

[1]Praxiling UMR 5267 – diwersy.sascha@univ-montp3.fr dodane.christelle@univ-montp3.fr Hirsch.fabrice@univ-montp3.fr luxardo.giancarlo@univ-montp3.fr
[2]Université Paris 8 – ivana.didirkova@univ-paris8.fr

Abstract
We present a corpus built upon the broadcast addresses of president Emmanuel Macron, delivered during the health crisis in 2020 and 2021. We integrate the orthographic transcriptions and the digitized audio recordings. This corpus is tokenized on the basis of both its text content and the related acoustic signal. The transcriptions then get a morphosyntactic tagging, a lemmatization and an annotation according to dependency relations. As a result, we make use of a linguistic environment organized according to a data model integrating acoustic, prosodic, (morpho)-syntactic and lexical features, thus allowing a processing both with textometric software, such as TXM, and with speech analysis software.

Keywords: textometry, written corpora, spoken corpora, exploratory data analysis.

Résumé
Nous avons construit un corpus constitué par les neuf allocutions télévisées du président Emmanuel Macron, prononcées pendant la crise sanitaire en 2020 et 2021, en intégrant les transcriptions orthographiques et les enregistrements audio numériques. Ce corpus est soumis à une segmentation textuelle mais aussi une segmentation à partir du signal acoustique. Successivement, les transcriptions ont fait l'objet d'un étiquetage morpho-syntaxique, d'une lemmatisation et d'une annotation au niveau des relations de dépendance. L'ensemble a été fusionné sous forme d'une base textuelle selon un modèle de données associant des traits acoustiques, prosodiques, (morpho-)syntaxiques et lexicaux, et permettant ainsi une exploitation aussi bien par des logiciels textométriques, comme TXM, que des outils d'analyse de la parole.

Mots clés : textométrie, corpus écrits, corpus oraux, analyse exploratoire des données

1. Introduction

L'analyse de discours politiques représente l'un des premiers domaines historiquement ciblés par la textométrie. Dans ce domaine, l'étude des évolutions chronologiques revêt un intérêt majeur. C'est ainsi qu'ont été forgées les notions de « série textuelle chronologique » pour caractériser des corpus textuels ordonnés selon le temps, respectant certains critères d'homogénéité, et de « temps lexical » pour décrire les variations (accroissements, disparitions ou discontinuités), observées dans le lexique (Salem, 1988). Afin d'étudier de tels corpus, les techniques d'analyse de données multivariée peuvent être utilisées : elles se fondent en principe sur une mesure de l'écart par rapport à une distribution régulière attendue. D'autres méthodes plus spécifiques ont été proposées, notamment des méthodes récursives permettant d'aboutir à une périodisation du corpus et à l'identification de champs lexicaux spécifiques. Dans ce domaine des séries textuelles chronologiques, rares sont les tentatives de traiter des corpus intégrant des données sur la parole ou l'interaction. Pourtant, les textes politiques résultent souvent de transcriptions de l'oral, qu'il s'agisse d'interventions monologales ou d'interactions verbales entre différents interlocuteurs. C'est la raison pour laquelle l'objectif de cette proposition est de prendre en compte différents paramètres phonétiques et prosodiques (les pauses notamment) en vue de mettre en évidence un « temps phonétique », dont les continuités ou discontinuités seraient éventuellement synchrones avec le temps lexical.

2. Outils et traitements

Afin d'intégrer les deux sources, les transcriptions (données écrites) et les enregistrements audio (données orales), en un corpus unique, nous avons effectué les traitements suivants :

1 Collecte des versions vidéo et audio des allocutions de Macron à partir de la chaîne *YouTube* de France 24 par l'intermédiaire du navigateur *Firefox* avec l'extension *YouTube Video and Audio Downloader*[1];

2 Collecte de la version transcrite des mêmes allocutions, telles que diffusées sur le site de l'Elysée via la page https://www.elysee.fr/toutes-les-actualites;

3 Traitement de la version audio au moyen du logiciel d'analyse de la parole Praat (Boersma et Weenink, 2021) pour la détection automatique d'unités interpausales et création d'une version TextGrid par l'alignement avec les transcriptions récoltées[2];

[1] https://addons.mozilla.org/fr/firefox/addon/youtube_downloader_webx/

[2] Les versions écrites diffusées par l'Elysée présentant quelques divergences, les transcriptions utilisées pour l'alignement ont fait l'objet d'une adaptation aux paroles réellement énoncées dans l'allocution retransmise à la télévision.

4 Conversion, par le logiciel *TEICONVERT* (Parisse *et al*., 2020), de la version TextGrid des allocutions au format XML TEI_CORPO[3], conforme au module TEI de transcription de la parole;
5 Enrichissement de la version XML par l'ajout d'annotations (morpho-) syntaxiques (parties du discours, traits morphologiques, relations de dépendances)[4] et de lemmes au moyen de l'analyseur *Stanza* (Qi *et al*., 2020);
6 Projection des transcriptions phonétiques incluant une délimitation syllabique répertoriées dans le dictionnaire électronique *GLAWI*[5] (Hathout & Sajous, 2016) aux formes issues de la tokenisation par Stanza;
7 Transformation des documents XML annotés au format VRT (verticalized text) servant d'entrée à l'indexation par CQP (Corpus Query Processor) (Evert, 2021);
8 Importation dans le logiciel de textométrie TXM dans le mode CQP.

Le corpus résultant, appelé *Covidis9* (Covid-19 : discours présidentiels) , est donc soumis à une annotation lexicale, syntaxique et phonétique et analysé selon les méthodes courantes de l'exploration textométrique, tout en y incluant des variables acoustiques telles que la durée des pauses ou le débit de parole. Les variables acoustiques sont intégrées à la fois aux annotations des unités lexicales dérivant de la tokenisation des textes et aux propriétés de structure déduites de la structuration XML.

3. Structuration et encodage du corpus

3.1. Modèle conceptuel

Le modèle conceptuel utilisé pour notre étude dérive de celui couramment défini en textométrie. Celle-ci structure un corpus en textes composés d'unités lexicales et associe des propriétés aussi bien aux unités lexicales qu'aux structures de plus haut niveau telles que les phrases, les textes ou le corpus dans sa globalité. Dans le cas de notre corpus d'étude, ce modèle se caractérise par l'ajout, à côté des unités de texte et de phrase, de deux unités de structure identifiées suite au traitement phonétique:

- la pause : un silence dans le signal acoustique,
- l'unité inter-pausale : une unité de la parole segmentée dans un énoncé et délimitée par des pauses

[3] https://ct3.ortolang.fr/teiconvert/
[4] Suivant le schéma Universal Dependencies (UD ; documenté, pour le français, à la page https://universaldependencies.org/fr/index.html)
[5] Téléchargeable à l'adresse http://redac.univ-tlse2.fr/lexiques/glawi.html.
[6] Le corpus est consultable en libre accès sur la plateforme ORTOLANG à l'adresse : https://www.ortolang.fr/market/corpora/covidis9

De par l'intégration des unités acoustiques en question, nous définissons deux propriétés de structures spécifiques : les propriétés colligationnelles[7] et les propriétés d'empan. Les propriétés colligationnelles permettent de décrire la position d'une unité de structure par rapport aux unités de structures de niveaux supérieurs dans la hiérarchie (dans l'ordre ascendant : pause ou unité inter-pausale – phrase – texte). Les propriétés d'empan caractérisent les unités de structure en termes de durée (mesurée en millisecondes), de vitesse d'articulation ou de débit (calculées en rapportant le nombre de syllabes des mots regroupés par une unité de structure à la durée de cette même unité) le cas échéant. Le tableau 1[8] donne les caractéristiques statistiques globales du corpus avec ses sous-échantillons.

3.2. Encodage des unités de structure et propriétés

Les concepts précédents sont encodés dans les documents résultants des différentes étapes de notre chaîne de traitement et, en dernier ressort, par le processus d'importation dans TXM. Dans l'ordre des traitements opérés, les unités de structure sont : (1) les allocutions (le texte) correspondant aux documents collectés, (2) les pauses et unités inter-pausales identifiées au moyen de Praat, (3) les phrases délimitées lors de la segmentation automatique des échantillons textuels opérée en amont de l'analyse syntaxique mise en œuvre par la chaîne d'annotation du logiciel Stanza. Pour les allocutions, nous avons retenu, à côté d'un identifiant posé par défaut, la date de diffusion. Les jeux de descripteurs associés aux phrases, unités inter-pausales et pauses sont conçus de manière comparable et incluent les traits colligationnels ainsi que les propriétés de durée, de vitesse d'articulation (unités inter-pausales) ou de débit (phrases). Afin de pouvoir être traitées par TXM (en tant que variables catégorielles), la durée, la vitesse d'articulation et le débit ont été représentés aussi bien par leur valeur brute (variable numérique) que par l'une des trois propriétés associées aux trois terciles résultant d'une subdivision en trois intervalles égaux. Les bornes d'intervalle qui délimitent les groupes de valeurs pour les différentes unités avec les propriétés concernées sont précisées dans le tableau 2. Le tableau 3 donne un aperçu de l'encodage des unités de structure du corpus avec leurs propriétés et les modalités associées. Pour ce qui est des propriétés associées aux

[7] Nous faisons appel à la notion de colligation textuelle, telle que proposée par Hoey (2005), qui décrit la préférence positionnelle d'une unité lexico-grammaticale dans un texte donné. Elle est ici généralisée à des unités caractérisées également par des traits phonétiques.

[8] Les tables et figures illustrant le présent article sont consultables à l'adresse suivante : https://www.ortolang.fr/market/corpora/covidis9?path=%2FJADT2022
(à noter que dans la première version du corpus, utilisée ici, la neuvième allocution n'a pas été incluse).

unités lexicales, celles-ci correspondent aux annotations obtenues au moyen de la chaîne de traitement Stanza (catégorie grammaticale, lemme, traits morphologiques et relation de dépendance syntaxique) avec, en extension, une réplication des propriétés de la tête syntaxique de chaque token et ses traits de colligation textuelle (au niveau du texte entier, des phrases et des unités inter-pause). En plus de leur modélisation comme unité de structure, les pauses ont également été associées à des propriétés au même niveau que les unités lexicales. Le modèle des propriétés lexicales du corpus est documenté par le tableau 4.

4. Résultats

Nous présentons d'abord deux traitements analytiques réalisés dans TXM sur notre corpus, l'un prenant en compte des propriétés d'unités lexicales, l'autre des propriétés d'empan (durée des pauses, débit au niveau des phrases). Nous réalisons deux Analyses Factorielles des Correspondances (AFC) sur les lemmes du corpus présents avec une fréquence supérieure ou égale à 4 :

- dans la première AFC (figure 1), par filtration de certains mots grammaticaux (ponctuation, déterminants, prépositions, numéraux, auxiliaires, conjonctions, pronoms, outre les tokens catégorisés 'X' correspondant aux pauses) en appliquant la requête : [upos!="(PUNCT|DET|ADP.*|NUM|AUX|.CONJ|PRON|X)"]. L'étude des résultats conduit à distinguer un regroupement des allocutions en quatre périodes (deux sur l'axe 1, deux sur l'axe 2).
- à partir des 25 lemmes fournissant les plus fortes contributions sur l'axe 1, une deuxième AFC (figure 2) accentue l'opposition entre les premières allocutions (mars 2020) et celles de sorties de crise (mars 2021 et juillet 2021 – le point-colonne 2021-07-12 fournissant la meilleure contribution relative sur le plan (1,2)). Par rapport à un ordonnancement « attendu » des différentes allocutions, la représentation obtenue fait ressortir ici la permutation entre les deux premières de mars 2020 (celle du 2020-03-16 se distinguant de toutes les autres) et la position « aberrante » de l'allocution du 2020-06-14.

L'analyse de spécificités illustrée par la figure 3 permet de visualiser la distribution des durées des pauses dans les différentes allocutions (les durées sont classées en trois catégories C1, C2 et C3, pour durées courtes, moyennes ou longues). On remarquera que les pauses longues sont surreprésentées dans les trois premières allocutions et qu'elles deviennent rares à partir d'octobre 2020. La distribution spécifique des pauses longues et courtes se reflète dans les différences très marquées du débit des phrases, qui font apparaître la même configuration mettant en opposition les allocutions de la première vague épidémique à celles à partir de la deuxième vague, avec une sur-représentation des phrases à débit ralenti (S_rate_C1) pour les unes et une sur-représentation des phrases à débit accélérée (S_rate_C3) pour les autres (figure 4). D'après Duez

(1999), les différences concernant la longueur des pauses ainsi que le débit et la vitesse d'articulation dans la parole politique peuvent être considérées comme étant des marques traduisant des situations au pouvoir différentes. Pauses brèves et débit accéléré indiquent la volonté d'optimiser l'exploitation du temps de parole pour accumuler arguments et idées exprimés – c'est le discours de la persuasion et de la (re)conquête du pouvoir. Pauses longues et débit ralenti, en revanche, reflètent la position des dominants, et participent en dernière analyse de la construction de l'ethos présidentiel – c'est le discours du pouvoir qui s'exerce. Dans le cas des allocutions de Macron, les différences observées quant aux traits acoustiques en question relèvent bien de préalables stratégiques divergents. Au début de la pandémie, face à ce qui allait devenir la première vague, Macron apparaît politiquement affaibli et en quête d'une légitimité largement perdue (cf. Sadoun-Kerber & Wahnich, 2022). Les premières allocutions vont être l'occasion de restaurer son autorité, restauration qui passe par la reconstruction de l'ethos présidentiel et l'activation de certains mythes constitutifs du mandat suprême de la Ve République, qui sont relayés par les thèmes marqués par les mots spécifiques de ces interventions télévisées en mars et avril 2020 : le mythe de l'homme providentiel, capable de passer à l'initiative face à la crise [2022-03-12: *prendre* (indice de spécificité : 3,5) et *mesure* (2,6), associés dans la locution verbale *prendre des mesures*, *réagir* (2,7)], mythe central modulé par celui du "Père" (Sadoun-Kerber & Wahnich 2022), garant du bien-être des concitoyens [2022-03-12 : *santé* (2,8)] ; le mythe du "Chef" (Sadoun-Kerber & Wahnich 2022) légitimé à responsabiliser ses concitoyens [2022-03-12 : *compter* (4,7) dans l'emploi avec COI marqué par *sur*] et à orchestrer le contrôle de leurs comportements [2022-03-16 : *consigne* (2,9) ; 2022-04-13 : *règle* (2,6)], voire celui du "chef des armées" (ibid.) [2020-03-12 : *mobilisation* (2,1) ; 2020-03-16 : *guerre* (6,2)]. A partir de la deuxième vague pandémique, d'abord sous-estimée pour son ampleur, Macron est confronté au risque de perdre le crédit supposément acquis grâce à sa gestion de crise, et il se voit obligé de persuader la population du bien-fondé de l'action gouvernementale. Notons entre autres que parmi les termes spécifiques de l'allocution du 28 octobre 2020, on trouve désormais à côte du nom *mesure* (2,9), qui renvoie en partie au confinement mis en place pendant la première vague [*printemps* (5,8)], celui de *stratégie* (2,7), qui souligne le caractère rationnel de son action. Cette même rationalité se voit d'ailleurs déclinée à travers un ensemble de séquences descriptives et argumentatives, cadrées de manière méta-discursive par une série de phrases interrogatives partielles marquées par le terme interrogatif *quel* (2,8). Pour confirmer que les oppositions entre allocutions que nous venons d'observer aussi bien sur le plan acoustique que sur le plan lexical ne relèvent pas d'un parallélisme aléatoire, il suffit de mettre en œuvre des requêtes qui ciblent des unités combinant les deux niveaux de traits. Si l'on interroge par exemple, au

sein d'une partition par allocution, la distribution des mots en conjonction avec le débit de la phrase dans laquelle ils sont employés, on retrouve non seulement le même regroupement de textes (cf. l'AFC en figure 5) qu'avec les calculs déployés séparément, mais aussi en grande partie les mêmes items spécifiques. Notons, pour terminer ce bref aperçu de résultats textométriques, que les différences de spécificités acoustiques observées se reflètent également de par la répartition au fil des allocutions des traits concernés, en l'occurrence le débit des phrases. Ainsi, on constate que le contraste du profil acoustique des différents groupes d'allocutions identifiés supra se confirme à travers les colligations textuelles des phrases à débit ralenti et accéléré. Les courbes de progression obtenues pour les allocutions du 2020-03-16 et du 2020-10-28 (figure 6), qui se distinguent par les distributions les plus opposées, montrent une répartition nettement différente des deux catégories de phrases selon les parties initiale, médiane et finale de chacune de ces allocutions, avec des pics locaux à des endroits bien différents. A côté d'une « dominante » acoustique, établie en termes de sur-emploi respectif des phrases à débit ralenti (76 occurrences, score de spécificité de 6 pour le 2020-03-16 – voir figure 4) ou accéléré (110 occurrences, score de spécificité de 14 pour le 2020-10-28), on voit donc apparaître, à l'égard des allocutions concernées, un cadençage caractéristique (une distribution plus régulière pour le débit « dominant »), dont l'analyse détaillée sera réservée à une future contribution.

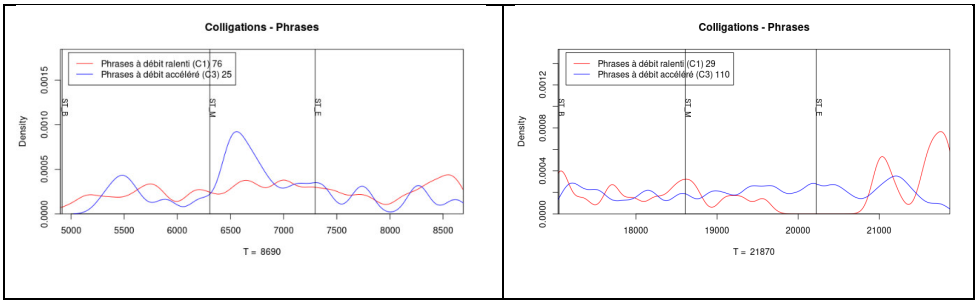


Figure 1. Progression des propriétés colligationnelles pour les allocutions du 2020-03-16 (gauche) et du 2020-10-18 (droite)

5. Conclusion

Avec cette étude, nous proposons donc de démontrer que les méthodes de l'analyse exploratoire des données mises en œuvre par le logiciel TXM peuvent s'appliquer non seulement à des unités lexicales et leurs propriétés morphosyntaxiques mais aussi à des traits phonétiques tels que les pauses, la vitesse d'articulation ou le débit de la parole et que la proximité des

classifications qui résultent des opérations appliquées fournissent des indices utiles dans l'interprétation. Parmi les extensions que nous envisageons à présent pour notre modèle, nous pouvons mentionner l'intégration de caractéristiques prosodiques.

Bibliographie

Boersma, Paul & Weenink, David (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.2.04, retrieved 18 December 2021 from http://www.praat.org/

Duez, Danielle (1999). La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de langues* 7(13), 91–97. https://doi.org/10.3406/flang.1999.1242.

Evert, Stefan & The CWB Development Team (2021). *The IMS Open Corpus Workbench* (CWB) - CQP Interface and Query Language Manual, retrieved 10 March 2022 from https://cwb.sourceforge.io/files/CQP_Tutorial.pdf.

Hathout, Nabil & Sajous, Franck (2016). Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. *Proceedings of the tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 136-1376, Portorož, Slovenia.

Hoey, Michael (2005). *Lexical priming: a new theory of words and language*. London ; New York: Routledge.

Parisse, Christophe, Etienne, Carole & Liégeois, Loïc. (2020). TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI, Journal of the Text Encoding Initiative. TEI Consortium. https://halshs.archives-ouvertes.fr/halshs-03043572.

Qi, Peng, Zhang, Yuhao, Zhang, Yuhui, Bolton, Jason & Manning, Christopher D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Sadoun-Kerber, Keren & Wahnich, Stéphane (2022). Emmanuel Macron face à la Covid-19 : un Président en quête de réparation d'image. *Argumentation et analyse du discours* 28. https://doi.org/10.4000/aad.6113. http://journals.openedition.org/aad/6113 (11 May, 2022).

Salem, André (1988). Approches du temps lexical, Mots 17, 105–143. https://doi.org/10.3406/mots.1988.1401.

The “Augmented” Historian

Alexandre Dniestrowski, Louis Allard*

Sorbonne Université Lettres Paris – alexandre@dniestrowski.net
louis.allard@etu.sorbonne-universite.fr

Abstract
Our research focuses on digital technologies applied to humanities and social sciences (DH) and more specifically to history. By taking advantage of innovative technologies (natural language processing, relational graph databases and web data) we explore different ways to enhance and extend the historian’s research and workflow capabilities and thus transform him or her into an “augmented” historian. The improved digital workflow is illustrated in a historical case study showing the steps necessary to undertake a correlation analysis to determine whether, in the later years of the Roman Republic, family relationships predominated among political factions.

Keywords: Digital Humanities - History – Late Ancient Rome Republic – Information Retrieval – Knowledge Management – Research Workflow - NLP – Relational Graph Database – Semantic Web

1. Introduction

The aim of historical research is to unearth the past in order to better understand the mechanisms of human societies.

For this lecture presentation, we take as an example the relationship between Octavian (the future emperor Augustus) and the consuls during the last years of the Roman Republic (43-30 BC). The consuls were the most important politicians in the organisation of Rome, and by analysing the origins, cultural and political backgrounds of these key witnesses, the historian gains a better understanding of the subtle changes of Roman society during this period.

Strengthened by this fine reconstitution of knowledge, the historian will be able to identify continuities or breaks between Late Roman Republic and High Roman Empire.

Our research on digital technologies applied to history aims to establish an operational framework for the historian that will improve his or her efficiency and enable new capabilities making him or her an “augmented” historian. This effort in the digital humanities is similar to many ongoing initiatives and

* Master/PhD at Sorbonne Université Lettres under supervision of Prof. Michèle Coltelloni Trannoy (UFR History)

platforms that are mentioned in (CUNY, 2022).
Before describing the main features of the enhanced digital workflow, we need to understand the obstacles that the historian must overcome in his or her scholarly activities.

2. Information Extraction and Knowledge Management Challenges

The historian follows a scientific process with a precise methodology (Nicolet, 1974), mainly:

1 Exploring Ancient Sources (*most time-consuming phase*):

a. The diversity of the literary corpus, the need to translate ancient languages and the search for documents available in digital format require an extremely long reading before an appropriate selection can be made

b. Other sources like literary, archaeological, epigraphic, numismatic materials also require careful consideration.

2 Historiographical literature (*history is a collective work*):

a. Compilation of publications (in various languages) related to the field of study.

b. Making the most of these historical works while keeping a critical eye on possible biases

3 Analysis (*added value contribution to collective historical work*):

a. Undertake scientific analysis following established best practice

b. Propose a new perspective or a synthesis with all appropriate justifications

Our historical study of Octavian contains a number of specific obstacles, such as:

- Source documents for our study period are limited due to the deliberate concealment for political purposes during the subsequent principate period.
- Early literary sources are fraught with ambiguities and translation problems
- Sources in the historiographical literature, such as "The Roman Revolution" (Syme, 1939), may be abundant on key figures, but very scarce on others that are important for our study
- Roman onomastics restrict first and last names: homonyms[1] must be disambiguated
- Sources do not always name a person in the same way[2]
- Time is a key factor that must also be represented in the knowledge repository

From these many obstacles, it is possible to infer certain requirements, such as:

[1] Such as Cornelius Messalla, Caius Cornelius Messalla Corvinus and Caius Cornelius Messalla Rufus

[2] Augustus, depending on the source can be called Caesar, Caesar the Younger or Octavian.

- A reading "accelerator" with information extraction and annotation editing
- A manageable repository to organize all the various information collected in an easily readable form with a powerful query mechanism to select subsets of the data
- Flexible statistical and analysis tools
- Because of the complexity of the historians' tasks, any 'solution' must be fully interactive
- Historians may not have extensive training in information technology; the digital solution must provide favourable operating conditions appropriate to its capacity

3. The "Augmented" Historian Workflow

The following diagram illustrates the overall structure of the workflow framework:

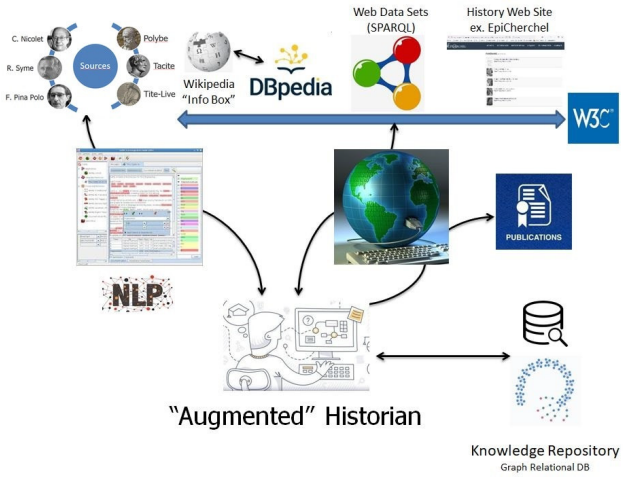


Figure 1. Digital Framework Architecture

To illustrate the workflow, we will follow the thread of a particular analysis where the historian wants to study the correlations between family relationships and political factions.
The overall workflow goes through the following stages:

3.1. Historical Sources Exploration

This phase requires a careful reading of many historical documents where, very often, there is no relevant information within the scope of the historical study. When something relevant is found the historian must extract the information and, if necessary, include additional annotations.
This very time-consuming manual process prompted us to look for a tool that

could at least ease the historian's burden in this task. We chose GATE (GATE, 2022) because it provides the historian with various NLP processors in a relatively easy to understand machine-like paradigm.
The corpus of documents selected for the study of Octavian was processed by GATE using specific gazetteers adapted to annotate Roman person names (Figure 2) and link each identified person to a "canonical name" in order to resolve homonymy issues. Where necessary, the historian will modify or add information (comments, text extractions) to the annotations.
Once the document analysis is complete, the annotation data is exported to a csv file for further processing or, if required, annotations from the internal Gate database can be saved in xml for further selective extractions.

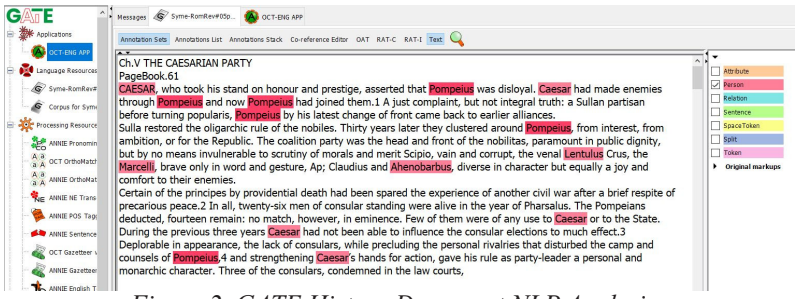


Figure 2. GATE History Document NLP Analysis

3.2. Building the Knowledge Database

At this stage, the historian has collected a great deal of information and now needs to organise all this knowledge and data into a repository that will allow him or her to undertake a selective cross-sectional analysis.
The Neo4j (Neo4j, 2022) relational graph database (Robinson & al., 2015) was selected primarily because its 'entity-relation-property' schema which lends itself well to the creation of a knowledge model that is comprehensible to the historian and sufficiently flexible to cope with the diversity of the data and to allow the historian to progressively build his or her knowledge database .
Since history is essentially evolutionary, the historian must represent time. The Octavian study spans a relatively short period (43-30 BC), the unit of time selected is one year.
As the Octavian study focuses on the family, social and political relationships between Octavian and the roman consular, all these elements are explicitly represented in the knowledge schema and, where appropriate, should be time-stamped to record the many developments.
For example, the node '*Gens*' contains information about a consular's surname network, while the 'Faction' node is related to membership of political groups that may change during the study period. Another factor is that family

relationships are largely a product of marriage. Apart from *Gens*, a static concept, all these entity-relationships are dynamic and evolve over the years of our historical study period.
All these elements are represented as nodes, relationships and properties as shown in figure-3.

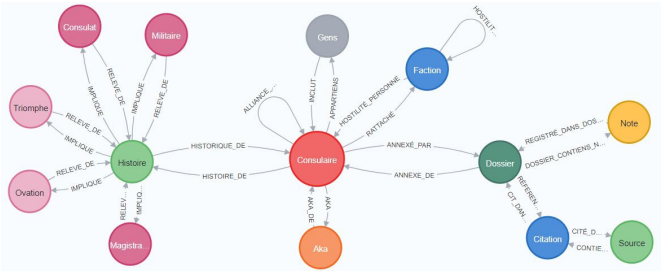


Figure 3. Octavian Knowledge Repository Schema (Neo4j)

Once a knowledge schema is defined, the database is loaded with all the data collected in the investigation phase (essentially csv files) and supplemented with ad hoc data, some of which are extracted from the Web (Antoniou & al., 2012) such as DBpedia datasets to complete properties such as family life data, honorary titles, judicial mandates or military charges : Augustus About page is a good example (DBpedia, 2022).
The historian can also create "notebook" databases from bulk data extractions (NLP or Web) imported using XML parsing filters and refined by the historian to add new complements to the knowledge base.
It's important to note that the workflow is not linear and there are many iterations between the exploration of sources and progressive establishment of the knowledge base.

3.3. Historical Analysis Investigations

The historian can now begin the cross-sectional analysis of the knowledge base in order to confirm initial assumptions or to identify new correlations.
For example, he wants to estimate the consistency between family relationships and political faction membership. In other words, is the family relationship the predominant factor in faction membership?
It would be very tedious and time consuming to do such a cross-sectional analysis directly from the csv data files created during the historical source exploration phase.
This is where the efforts to build the knowledge database is rewarded.
Using a short and relatively simple query script, written in a database query language (Cypher in our case), relevant data will be collected and organised into

data tables containing the key elements needed for the analysis.
These tables will then be transferred to Excel for further statistical or logical processing.
In this way, the historian will progressively advance in his knowledge and, step by step, he will be able to draw his conclusions.
As we can see from the description of the workflow, the process is iterative, the automation is selective and the historian is in full control.

4. Preliminary Conclusions

The Octavian study shows encouraging results in the application of digital technologies to improve productivity and capabilities of the historian.
The "Augmented" Historian workflow benefits mainly from: faster reading of historical documentation, powerful selection of knowledge elements from the graph relational database, flexible processing (statistics or logical) and multiple presentations of analysis results.
The workflow described in this paper, can be further improved through increased automation, for example simplifying the transformation of bulk data downloads (notebook databases) into knowledge items that are stored directly in the knowledge database.
Another promising avenue, is to extend usage of NLP (Maynard & al., 2016) to mobilise more sophisticated techniques such as special rule-based processing and machine learning.
These new perspectives (improved nlp processing and workflow automation) will be at the heart of our future research to further 'augment' the capabilities of historians.

References

Antoniou G., van Harmelen F. and Primer A. S. W. (2004). MIT Press. *Cambridge*, *MA*.
CUNY. (2022). Cuny Academics Digital Humanities Resource Guide. Retrieved 03 07, 2022, from Cuny Academic Commons:
https://wiki.commons.gc.cuny.edu/Sample_Projects/
DBpedia. (2022). *DBpedia*: https://dbpedia.org/page/Augustus
GATE. (2022). *GATE*: https://gate.ac.uk/
Maynard D., Bontcheva K. and Augenstein I. (2016). Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology,* 6(2), 1-194.
Neo4j. (2022). *Neo4j*: https://neo4j.com/
Nicolet C. (1974). L'ordre équestre à l'époque républicaine (312-43 av. J.-C.). *Tome 1 : Définitions juridiques et structures sociales*. Paris : Éditions E. de Boccard.
Robinson I., Webber J. and Eifrem E. (2015). *Graph databases: new opportunities for connected data*. "O'Reilly Media, Inc.".
Syme R. (1939). *The Roman Revolution*. Oxford University Press.

Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle

Chuanming Dong[1], Philippe Gambette[2], Catherine Dominguès[3]

[1]LASTIG, Univ Gustave Eiffel, ENSG, IGN et ADEME – chuanming.dong@ign.fr
[2]Univ Gustave Eiffel, CNRS, LIGM – philippe.gambette@univ-eiffel.fr
[3]LASTIG, Univ Gustave Eiffel, ENSG, IGN – catherine.domingues@ign.fr

Abstract
We seek to characterize events related to industrial pollution, which are extracted from administrative texts. Each event can be characterized by a trigger that reflects its nature (administration, depollution, etc.). We propose a typology to categorize these triggers and a hybrid method to automatically perform this categorization. This method combines a machine learning approach (based on word embeddings and supervised learning) with a textometric approach (based on the visualization of the co-occurrence of frequent words). It allows characterizing event triggers automatically with an f-score of 80%.

Keywords: statistical analysis, natural language processing, machine learning, clustering, tree cloud, textometry

Resumè
Nous cherchons à caractériser des événements liés à la pollution industrielle, extraits de textes administratifs. Chaque événement peut être caractérisé par un déclencheur qui rend compte de sa nature (administrative, de dépollution, etc.). Nous proposons une typologie pour catégoriser ces déclencheurs et une méthode hybride pour réaliser automatiquement cette catégorisation. Cette méthode combine une approche d'apprentissage automatique (fondée sur des plongements de mots et un apprentissage supervisé) avec une approche textométrique (fondée sur la visualisation de la cooccurrence des mots fréquents). Elle permet de caractériser les déclencheurs d'événements automatiquement avec une f-mesure de 80%.

Mots clés : analyse statistique, traitement automatique des langues, apprentissage automatique, clustering, nuage arboré, textométrie

1. Introduction

La pollution est une des préoccupations centrales des français, et les activités industrielles constituent une source potentielle et importante de pollution. En

France, c'est le ministère en charge de l'Écologie qui collationne et diffuse les informations concernant les activités industrielles et les sites sur lesquels elles ont été ou sont exercées, par l'intermédiaire des bases de données BASOL, BASIAS et S3IC. En particulier, BASOL inventorie les sites et sols pollués, ou potentiellement pollués, appelant une action des pouvoirs publics, à titre préventif ou curatif.

Dans le cadre normal de leurs activités, les industries manipulent ou produisent des produits éventuellement toxiques. Pour chaque entreprise, BASOL rassemble dans différents champs les informations concernant l'activité industrielle et les produits manipulés. Cependant, des événements peuvent aussi se produire qui ne concernent pas l'activité industrielle normale et peuvent être à l'origine de pollutions dues par exemple à une quantité anormale de certaines substances ou à leur contact accidentel. Ces événements sont exceptionnels par rapport à l'activité industrielle normale, et à ce titre non associés à des champs de la base; ils sont alors décrits de manière textuelle et ces textes figurent dans la base, contribuent à la description de l'entreprise mais, non structurés, ne sont pas interrogeables par l'intermédiaire de l'interface d'interrogation et de visualisation fournie sur le site officiel. Pourtant, ces textes contiennent des informations cruciales pour évaluer la situation des sites industriels, en particulier pour connaître les pollutions déjà constatées ou prévisibles ainsi que les actions de dépollution déjà mises en œuvre ou à prévoir. Ces informations intéressent, à divers titres, à la fois les citoyens et les experts. Nous visons donc à construire une mémoire des sites industriels, sous la forme d'une base de données unique (BDU) décrivant les activités normales et les événements exceptionnels concernant les sites industriels, en appariant et réorganisant les informations structurées contenues dans plusieurs bases de données, (dont BASOL) et en extrayant, des textes descriptifs contenus dans les mêmes bases, les caractéristiques des événements survenus sur les sites industriels.

Les éléments permettant de caractériser un événement industriel (noyau, activité, entreprise, polluant, etc.) ainsi que la méthode d'extraction de ces éléments dans les textes descriptifs ont été décrits dans (Dong *et al*., 2021). L'examen de ces caractéristiques a montré la diversité des événements potentiels. D'autre part, la restitution au public du contenu de la BDU, sous une forme la plus complète possible mais aussi intelligible, nécessite de catégoriser et hiérarchiser les informations. En particulier, tous les événements n'ont pas la même importance quant à l'évaluation de la pollution d'un site. L'objet de cet article est de présenter une typologie des événements industriels et les outils qui permettent de classer dans cette typologie, de la manière la plus automatique possible, les événements industriels identifiés dans les textes descriptifs. La démarche proposée s'appuie sur un corpus textuel de descriptions d'événements industriels issues de l'extraction de (Dong *et al*. 2021) et la caractérisation d'un événement pertinente

pour cette thématique. D'une part, elle propose une typologie des événements permettant de classer leurs noyaux en plusieurs classes sémantiques. D'autre part, elle combine une approche textométrique et des méthodes et outils d'apprentissage automatique pour classer dans cette typologie des mots caractéristiques de chaque classe, extraits de ces noyaux, que nous appelons « déclencheurs », et pour enrichir cet ensemble de déclencheurs. La mise en forme des données (préparation de la base de données BASOL et caractéristiques des événements industriels) est rappelée dans la section suivante. La construction de la typologie de ces événements est expliquée dans la section 3. La section 4 expose la démarche de classification des événements industriels dans cette typologie et d'augmentation du vocabulaire désignant les événements. Une conclusion ainsi que des limites de la démarche sont présentées dans la dernière section.

2. Mise en forme des données

Définie dans les années 1990 et renseignée par les directions régionales de l'Environnement, de l'Aménagement et du Logement, BASOL est une base de données publique qui vise à recueillir les informations sur les sites industriels potentiellement pollués. Chaque enregistrement correspondant à un site contient, outre des données structurées caractérisant l'entreprise et son activité, un texte qui décrit des événements relatifs à l'activité normale, à des incidents et des accidents conduisant éventuellement à des pollutions, à l'aménagement ou au réaménagement du site, etc. En 2020, BASOL contenait les informations de plus de 7 000 sites industriels en France. Les textes descriptifs de BASOL constituent le corpus de travail pour l'identification et l'analyse des événements industriels. Il est caractérisé par une langue soutenue et l'emploi d'un vocabulaire spécifique de ces événements industriels, et contient 155 587 phrases, soit 48 032 mots, dont des exemples suivent :

Un diagnostic initial et une évaluation simplifiée des risques (ESR) <u>ont été prescrites</u> par arrêté du 17 avril 2000 conformément à la circulaire du 3 avril 1996.

Les terres polluées correspondant aux zones P9S21 et PY19B2 <u>ont été excavées et évacuées</u> vers des installations de stockage de déchets et des centres de traitement biologique des terres en 2001 et 2002.

Un événement industriel est défini comme une action qui a eu lieu et a modifié la situation de l'entreprise. Les exemples précédents montrent que l'événement est caractérisé par un noyau (<u>souligné</u>) qui définit le type de l'événement, administratif (avec le déclencheur *prescrit*) dans le premier exemple et de réhabilitation (avec les déclencheurs excavées et *évacuées*) dans le deuxième. On s'intéresse dans cet article à un corpus constitué par les noyaux d'événements. Ce corpus a été construit en annotant automatiquement le corpus BASOL à l'aide

du modèle d'annotation automatique entraîné sur une portion de ce même corpus (Dong *et al*., 2021).

3. Caractérisation et typologie des déclencheurs d'événements

La diversité des déclencheurs ainsi que leurs conséquences différenciées sur la situation de pollution d'un site industriel nécessitent de construire une typologie de ces déclencheurs qui permettra de les regrouper sur des critères sémantiques et facilitera l'interrogation de la BDU. Afin de construire cette typologie, une méthode fondée sur la construction d'un nuage arboré des noyaux d'événements (Gambette et Véronis, 2009) a été mise en place. Elle permet de faire apparaître les mots les plus fréquents au sein de ces noyaux, organisés dans un arbre qui rapproche les mots apparaissant fréquemment proches l'un de l'autre au sein d'un noyau. L'observation de ce nuage arboré[1] permet d'identifier six classes d'événements qui constituent le point de départ de la typologie :

- **admin** : les actions provenant de l'autorité administrative (autorisation, demande, prescription, etc.) ;
- **nuis** : les actions décrivant des nuisances contre l'environnement ;
- **diag** : les actions techniques concernant le diagnostic de pollution ;
- **depol** : les traitements relatifs à la réhabilitation ;
- **indus** : le parcours administratif de l'entreprise (implantation, changement ou cessation d'activité, fermeture, etc.) ;
- **neutre** regroupe les actions qui ne peuvent pas être classées dans les classes précédentes, en particulier celles qui, comme : réaliser, effectuer, mettre en œuvre, opérer etc. annoncent la réalisation d'une autre action qui pourra être classée dans la typologie.

On peut donc créer une liste de mots simples spécifiques à chaque classe pour obtenir un premier ensemble de déclencheurs. Contrairement à une méthode d'apprentissage supervisé qui nécessite de connaître préalablement la liste des étiquettes, l'utilisation du nuage arboré permet d'identifier les étiquettes pertinentes. Cette méthode fondée sur les mots fréquents et leur cooccurrence est particulièrement adaptée dans notre situation où le sens des noyaux d'événements est fortement lié à leur contenu lexical.

4. Enrichissement des déclencheurs des noyaux

La capacité de représentation des nuages arborés est limitée à une centaine de mots les plus fréquents. Ceci ne permet de repérer qu'un nombre limité de déclencheurs, et pénalise les déclencheurs moins fréquents, et comme on peut

[1] voir ce nuage arboré dans le matériel supplémentaire de cet article, incluant le code Python développé et les données : https://github.com/DongChuanming/JADT2022_classification_et_enrichissement

l'observer avec la catégorie nuis, les catégories les moins fréquentes de la typologie (un problème également présent dans Gambette *et al*., 2018). Nous combinons donc deux approches pour augmenter le vocabulaire des déclencheurs d'événements classés dans la typologie. D'une part, une démarche itérative consiste à répéter l'extraction manuelle de mots depuis un nuage arboré représentant les noyaux d'événements ne contenant aucun des mots extraits auparavant. D'autre part, une démarche d'apprentissage automatique, à l'aide d'un algorithme de classification fondé sur des plongements lexicaux de mots extraits de ces nuages arborés, est adoptée pour capter également des termes peu fréquents.

4.1. Protocole d'enrichissement à partir des mots fréquents

Le peuplement de la typologie comporte deux axes : reconnaître les formes fléchies des déclencheurs déjà catégorisés (enrichissement horizontal), et reconnaître les nouveaux termes qui n'ont pas le même lemme que les déclencheurs déjà catégorisés (enrichissement vertical).

Le procédure de peuplement est itérative et chaque itération comporte deux étapes. La première étape, déjà mentionnée dans la section 3, vise à réaliser l'enrichissement vertical et consiste à obtenir un sac de mots (S1) en extrayant les déclencheurs manuellement à partir du nuage arboré dans les catégories de typologie. À la deuxième étape (pour réaliser l'enrichissement horizontal), une liste (L1) est générée pour chaque classe en ajoutant certaines formes dérivées manuellement (nom prédicatif si le verbe est un déclencheur, verbe à l'infinitif si le nom prédicatif est un déclencheur) des déclencheurs de cette classe, ainsi que les formes fléchies de tous ces termes, générées automatiquement à partir du dictionnaire DELA (Paumier, 2011). À l'itération suivante, un nouveau nuage arboré est construit à partir des noyaux d'événements ne contenant aucun déclencheur de la liste S1+L1. De même qu'à l'itération précédente, un nouveau sac de mots (S2) et les listes correspondantes de leurs formes dérivées et fléchies (L2) sont générés. Le processus itératif se termine quand le nombre de déclencheurs ajoutés n'augmente plus significativement.

Cette méthode permet donc non seulement de capter des termes un peu moins fréquents dans une démarche manuelle mais aussi de tester et d'alimenter une procédure d'enrichissement automatique à partir d'un classifieur, tout en priorisant l'enrichissement des classes qui contiennent le moins de termes.

4.2. Protocole d'enrichissement à partir d'un classifieur

Simultanément avec l'enrichissement manuel à partir des nuages arborés, une approche d'apprentissage automatique est mise en œuvre pour automatiser la procédure de classification des déclencheurs et finalement enrichir la liste des déclencheurs. À la fin de la n-ième itération, un classifieur est entraîné sur

l'ensemble (S1+...+Sn)+(L1+...+Ln) des déclencheurs déjà catégorisés, et est évalué sur l'ensemble Sn+1+Ln+1 des déclencheurs catégorisés à l'itération suivante. Les données d'apprentissage sont donc accumulées à chaque itération. En combinant les sacs de mots et les listes des formes fléchies, le classifieur est entraîné et évalué à la fois sur les enrichissements horizontal et vertical. La classification supervisée est fondée sur le perceptron multicouche (Haykin, 1994), et les plongements de mots sont fournis par le modèle de langue française CamemBERT (Martin *et al*., 2020).
Finalement, l'implémentation d'une méthode itérative permet de mieux cibler les efforts : ajuster le corpus d'apprentissage en ajoutant au fur et à mesure les déclencheurs les moins fréquents, raffiner le classifieur en intégrant les plongements de ces nouveaux déclencheurs. Ainsi, le classifieur obtenu à la fin de la première itération obtient, pour la classe nuis, une précision de 0,77 mais seulement 0,30 pour le rappel. En conséquence, dans l'apprentissage de la 2ème itération, le nombre des lemmes déclencheurs de la classe nuis est augmenté par rapport à ceux des autres classes, pour renforcer la flexibilité du classifieur dans la reconnaissance de cette classe. Ainsi le modèle est amélioré, obtenant à la fin de l'itération 2, une précision de 0,97 et une rappel de 0,70 pour cette même classe. Pour l'ensemble des classes, la f-mesure passe de 0,51 à la fin de la première itération à 0,80 à la fin de la deuxième[2].

5. Conclusion

Dans cet article, nous avons présenté une approche pour construire et enrichir une typologie sémantique concernant les actions de déclencheurs dans les événements industriels, qui combine les méthodes textométriques et statistiques. En combinant l'apprentissage automatique avec la classification manuelle itérative, une méthode d'apprentissage supervisé itératif est conçue, et celle-ci permet, en se fondant d'abord sur le vocabulaire le plus fréquent dans chaque classe, de créer rapidement des données d'apprentissage de grande taille, susceptibles d'être enrichies en utilisant un modèle de langue fondé sur des plongements. Un atout du résultat obtenu est que chaque classe thématique d'intérêt est caractérisée par un ensemble de déclencheurs susceptibles d'être présentés à une personne non experte du domaine pour l'aider à interroger et explorer la BDU. En revanche, cette approche n'a pas pour objectif d'optimiser à la fois la précision et le rappel de la classification des noyaux, qui permettrait des requêtes ciblées sur certaines classes plus précises de noyaux ou de

[2] détails des évaluations et liste de déclencheurs enrichie par la méthode itérative et le classifieur disponibles ici : https://github.com/DongChuanming/JADT2022_classification_et_enrichissement

déclencheurs, avec garantie d'exhaustivité des résultats. Le ciblage des déclencheurs les plus fréquents pour constituer le corpus d'apprentissage permet d'assurer un niveau minimum de représentativité d'apprentissage, les déclencheurs moins fréquents, absents du corpus d'apprentissage, pouvant être repérés par le calcul de plongements. Une limite de la découverte et du classement des déclencheurs provient de la segmentation en tokens (mots ou portions de mots) pour le calcul des plongements, peu propice à l'étiquetage d'expressions polylexicales.

Ultérieurement, le classifieur sera utilisé pour classer des déclencheurs déjà extraits du corpus BASOL. Le principal problème à régler est le cas où un noyau d'événement contient plusieurs déclencheurs de catégories différentes, comme dans les noyaux stopper l'extension de la pollution (catégories dépol et nuis) et avait prescrit un audit sécurité (catégories admin et dépol). Plusieurs pistes sont envisagées, par exemple établir un ordre de priorité entre les catégories. Finalement, les déclencheurs d'événement, caractérisés par la typologie, pourront être enregistrés dans la BDU avec leurs composants extraits du corpus BASOL.

Remerciements

Ce travail est en partie financé par l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME), 20, avenue du Grésillé- BP 90406 49004 Angers Cedex 01 France.

Bibliographie

Dong C., Gambette P. et Dominguès C. (2021). Extracting Event-related Information from a Corpus Regarding Soil Industrial Pollution. In Proc. of KDIR 2021, vol. 1, pp. 217-224.

Gambette P., Kyriacopoulou T., Lechevrel N. et Martineau C. (2018). *Anatomie, animaux, vocabulaire de la vivisection. Construire des ressources lexicales pour visualiser une thématique dans un corpus littéraire*. In Animalhumanité. Expérimentation et fiction : l'animalité au cœur du vivant, p. 223-232.

Gambette P. et Véronis, J. (2010). Visualising a Text with a Tree Cloud. In Proc. of IFCS 2009, Studies in *Classification, Data Analysis, and Knowledge Organization* 40, pp. 561-570.

Haykin, S., & Lippmann, R. (1994). Neural networks, a comprehensive foundation. *International journal of neural systems*, 5(4), 363-364.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D. et Sagot B. (2020). CamemBERT: a tasty French language model. In Proceedings of the *58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203-7219.

Paumier S. (2011). *Unitex - Manuel d'utilisation*, https://hal.archives-ouvertes.fr/hal-00639621/document

Automatic retrieving of derived Quechua verbs

Maximiliano Duran

Université de Franche-Comté, CRIT – duran_maximiliano@yahoo.fr

Abstract
In this paper, I present a method allowing the automatic retrieval of derived and inflected Quechua verbs embedded in a corpus. It consists of the formal construction of paradigms and grammars managing the grammatical combinations of verbal suffixes, I use these formalized grammars in order to extract all the verb forms contained in the corpus. Then, joining our Electronic Dictionary of verbs to some elementary statistical tools to these results, I have obtained a practical system that allows answering the following socio-lingüistic questions: a. how frequent is the use of multi-suffixation of verbs (derived by a combination of two or more suffixes) in everyday life communication? b. Is there a notable difference in this phenomenon between written and oral communication?

Keywords: Quechua verb derivation , Quechua, verbal suffixes

1. Introduction

Quechua is the Incas' language, spoken by over 6 million people in Peru, Equator, Bolivia, and Argentina. Quechua has an important characteristic: its verbal morphology is very powerful and highly regular. In this paper, we present Quechua's inflectional and derivational systems, and how we have formalized them by constructing NooJ grammars . With the help of these grammars and the Electronic Dictionary of verbs, that we have written by translating into Quechua the dictionary of Dubois & Dubois-Charlier, 1997, 2007. *Les verbes français* (LVF) NooJ's parser, see Silberztein, 2016, can automatically recognize any occurrence of verbal inflected form and analyze it. And, reciprocally, they may generate any inflected and derived form of a simple verb[1].

First, I explain how I formalized the conjugation of any Quechua verb. The key for this formalization is based on the structure of the set of Indefinite Tense endings (IT)[2] of Quechua and its morpho-syntactic behavior. This IT set is used in the conjugation of any verb in the indefinite tense[3]. Then, using the results of

[1] A verb form which does not contain any suffix.
[2] The indefinite tense suffixes are IT = (*ni, nki, n, nchik, niku, nkichik, nku*).
[3] We call « indefinite » tense (IT), because they serve to conjugate the Present: *rima-ni* [I talk], the Past: *rima-ni* [I talked], and in the Habitual mood: *rima-ni* [I am used to talk] at the same time.

Duran, 2017, who studied the different transformations of IT, I constructed the paradigms allowing to obtain all the nine sets of transformed endings (TRITn, n=1..9)[4]. This allowed Duran, 2017 to obtain a more comprehensive grammar serving to obtain the conjugated forms of multi-suffixed verbs, in all tenses and moods. (eg. *rimani*/ I talk, *rimaiman*/ I would talk, …)
Using the NooJ linguistic platform, developed by Sylberztein, 2016, I apply these formalized grammars to two corpora: one written and one oral. The written one is made of eight novels and some historical narrations in Quechua, to retrieve the verb forms bearing concordances with these grammars, and extract them. I applied then some simple statistic calculations on the set of frequencies. Finally, based on these results, I was able to answer the posed interrogations concerning the frequency of the use of multi-suffixation in the everyday life communication of the Quechua population.

2. Fundamental conjugation

The fundamental conjugation of a Quechua verb, the FC-structure, has the following form: Vroot+ITending, shown in Table 1:

Table 1. The indefinite-tense fundamental conjugation FC of a Quechua Verb.

QU Pronoun	*EN Pronoun*	*Verb root*	*IT Ending*
Ñoqa	I	root	*ni*
qam	You	root	*nki*
pay	he, she	root	*n*
ñoqanchik	we (inc.)	root	*nchik*
ñoqaiku	we (excl.)	root	*niku*
qamkuna	you	root	*nkichik*
paykuna	they	root	*nku*

In this table, we can see that each suffix marks a person in an exclusive way. Hence, for instance, if we see *miku**ni*** , I can say: it's the verb *mikuy* [to eat] conjugated in the first person, because of the presence of the suffix « *ni* ».
We formalize this conjugation structure by the following NooJ grammar:
IT = <B> (*ni*/IT+1+s | *nki*/IT+2+s | *n*/IT+3+s | *nchik*/IT+1+pin | *niku*/IT+1+pex | *nkichik*/IT+2+p | *nku*/IT+3+p) ; where we have:

[4] e.g. TRIT1 = (*i, nki, n, nchik, iku, nkichik, nku*) ; TRIT2 = (*i, iki, n, nchik, iku, ikichik, nku*),

1+s: first-person singular ; 2+s: second-person singular ; 3+s: third-person singular ; 1+pin: inclusif first-person plural; 1+pex: exclusif first-person plural; 2+p: second-person plural; 3+p: third person plural

For the future, the corresponding set of endings is Fending= (*saq, nki, nqa, sunchik, saqku, nkichik, nqaku*). Here also, we remark that each person is marked separately. We symbolize by FUT the corresponding formalized grammar of the future tens :

FUT = <B> (saq/F+1+s | *nki*/F+2+s | *nqa*/F+3+s | *saqku*/F+1+pex | s*unchik*/F+1+pin | *nkichik*/F+2+p | *nqaku*/F+3+p) ;

A more general typical Quechua conjugated verb form[5] has the following structure:

Vroot + IPS + ITending or Vroot + IPS + Fending

where: V: stands for the root of the verb

IPS[6]: Interposed suffix is the set of 27 derivations and 8 inflection suffixes

ITending: is the set of seven indefinite tense endings cited above

Fending: is the set of seven indefinite tense endings cited above.

According to the Quechua verb morphology, these suffixes, agglutinated to the root will generate 27 derived verbs, which may be conjugated. The formalized grammar of the corresponding paradigm[7] is shown hereafter:

V_TR_SIP1_CONJ0 = :PR |:F |:IP |:RQA_PR_PLU |:RQU_PR_PLU |:SPA |:SPA_PRM2_PLU |:RU_PR_PLU|:RA_PR_PLU | :SQA_PRM2_PLU | :PTI_PRM2_PLU |:WA_PRM5 |:V_SIP0_STIN | :V_SIP1_STIN | <B>(V_SIP1_INF)(ni/PR+1+s |nki/PR+2+s |n/PR+3+s |nchiks/PR+1+pin |ni/PR+1+pex |nkichik/PR+2+p |nku/PR+3+p |saq/F+1+s| nki/F+2+s | nqa/F+3+s | saqku/F+1+pex | nqaku/F+3+p | sunchik/F+1+pin | nkichik/F+2+p); the details of this construction can be found in (Duran 2017).

Applying it to the verb *llamkay* [to work], NooJ automatically produces all its conjugated forms in the Present, Past, Gerund, Habitual mood and Future tenses.

[5] In fact, a more general conjugated form should include the post-position suffixes (placed after the IPSendings or Fendings), PPS, studied by Duran, 2013, 2017. The presented paradigms serve as a basis for the more general case.

[6] IPS =(chaku, chi, chka, ykacha, ykachi, ykamu, ymana, ykapu, ykari, yku, ysi, kacha, kamu, kapu, ku, lla, mu, naya, pa, paya, pu, raya, ri, rpari, rqu, ru, tamu)

[7] The details of the construction of this paradigm can be found in Duran, 2017.

```
# NooJ V7 # Dictionary #Language is: qu
llamkani,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+1+s
llamkanki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+2+s
llamkanku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+3+p
llamkan,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+3+s
llamkankichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PR+2+p
llamkanki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+2+s+F
llamkanqa,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+3+s
llamkanqaku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+3+p
llamkasaq,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+1+s
llamkankichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+F+2+p
llamkai,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+2+s
llamkachun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+3+s
llamkasun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+p+1
llamkaichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+IP+2+p
llamkarqun,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PAPT+3+s
llamkarqunkichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+PAPT+2+p
llamkaspa,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER
llamkaspaiki,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+2+s
llamkaspanku,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+3+p
llamkaspai,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+1+s
llamkaspan,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+3+s
llamkaspaikichik,llamkay,V+FR="travailler"+FLX=V_TR_SIP1_CONJ0+GER1+2+p
```

NooJ: Success

Dictionary C:\Users\Toshiba\Documents\NooJ\qu\Lexical Analysis\z-flx.dic has been successfully compiled:
329 inflected forms.

OK

Figure 1. Automatically generated Conjugated forms in diferent tenses and moods for llamkay.

Do all of the automatically generated forms have an actual meaning? We answer this question by the affirmative in a separate work in progress.

3. Retrieval of inflected and derived verbs

Using the retrieval operation: Text – Locate <V> operators, we have found 2484 verb forms contained in the written corpus. This can be seen in Figure 2 (see the column labeled as Seq.).

Next, we looked for the verb forms bearing one derivation suffix. To get them, we have utilized the paradigm V_TR_SIP1_CONJ0, presented above, combined with two additional ones V_ITR_SIP1_CONJ0 and V_IMP_SIP1_CONJ0[8] which serve to locate the forms containing a intransitive and impersonal verb roots ; these paradigms include also the derivations by one IPS suffix followed by their corresponding inflections.

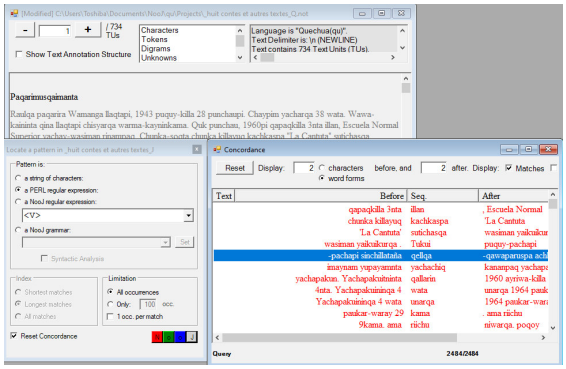


Figure 2. The totality of conjugated verb forms recognized by the gramma in the written corpus

[8] The detailed description of these paradigms, corresponding to intransitive and impersonal verbs, can can be found in Duran, 2017.

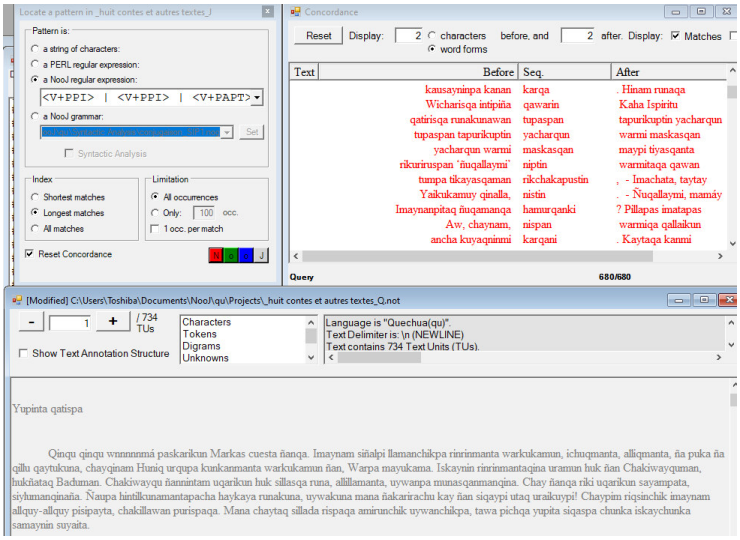


Figure 3. Verb forms containing one derivation suffix.

4. A socio-linguistic application

How frequent is the use of derived verbs in everyday communication ?

a. Analysing our first results, we remark that the mono-suffixed verb forms appearing in Figure 3, cover 27,38% of the totality of verb forms enbedded in the text.

b. For the **written** narrations, like our corpus of tales, a fequency analysis seems to show the numeric relevance of the use of he verb *niy*[to say] and its derivations: 114/680 (16,76%), followed by the verb *kay*[to be] and its derivations: 60/680 (8,82%). This fact contrasts with

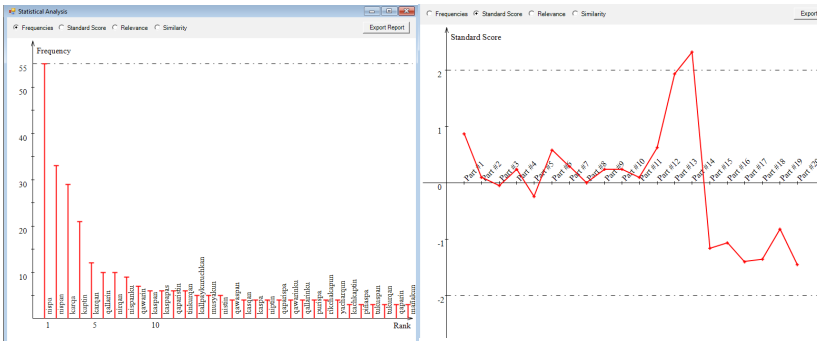


Figure 4 : Frequency of the use of the verb niy[to say] in written corpus and its standard score.

the values obtained for the case of the oral discourse , in which we found that the most frequently used verb is *kay*[to be] and its derivations (10/50) 20,00%, whereas the second most used one is the verb *niy*[to say] and its derivations (7/50) 12,72%.

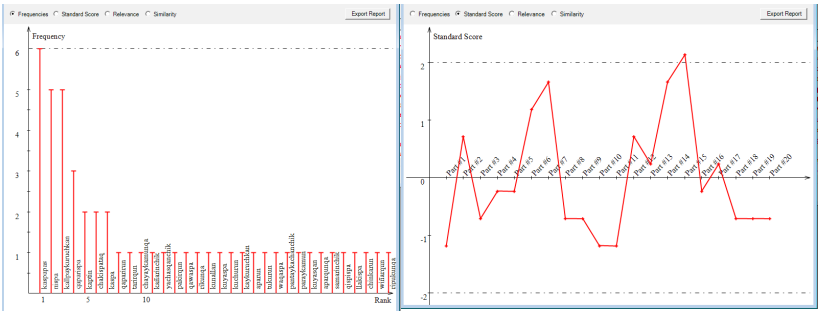


Figure 5 : Frequency and standard score of the verb kay[to be] in an oral discourse.

On the other hand, we must note that the derivation of a Quechua verb may be obtained also by combinations of two or more IPS suffixes. We may ask ourselves, how frequently does a speaker utilize this kind of complex verb form. Does the frequency of these multi-suffixed verbs dependent on the context of a written or an oral discourse[9]. Part of the answer to these questions may be foreseen in the following data.

Case of the **oral** discourse:

Total of verb forms, in the oral corpus, with more than 1 derivation suffix: 50

forms containing 2 or more derivation suffixes	Freq: 13	26%
forms containing 3 or more derivation suffixes	Freq: 6	12%

Case of the **written** discourse:

Total of verb forms, in the written corpus, with more than 1 derivation suffix: 680

forms containing 2 or more derivation suffixes	Freq: 60	8,8%
forms containing 3 or more derivation suffixes	Freq: 2	0,3%

[9] A transcription of a transmission of the program « *Musuq kallpa* » in radio Kispillacta (*Mikuy-kunamanta* 19/03/2003) was realized by C. Itier, 2003.

These values suggest that the utilisation of verb forms containing two or more derivation suffixes is more frequent in an oral discourse than in a written one. If this is thecase, it may be explained parhaps by the fact that Quechua remains, dominantly, an oral language.
Of course, this conconlusion, taken from these limited corpora, may not represent the general case, it should be verified with larger corpora. We hope that, the method and the paradigms we have presented in this work may contribute to do it.

5. Conclusion and perspectives

I have presented a method allowing the automatic retrieval of derived and inflected Quechua verbs embedded in a corpus. It consists in the construction of paradigms and grammars, in the NooJ formalism, managing the grammatical combinations of verbal suffixes. Using these formalized grammars, I have retrieved all the verb forms contained in a written and an oral corpus.
Joining our Electronic Dictionary of verbs[10] to some elementary statistical tools, I have obtained a practical system that help us to answer the socio-lingüistic questions. For the first question: How frequent is the use of multi-suffixation of verbs (derived by a combination of two or more suffixes) in everyday life communication? We have found that more than one fourth of the verb forms (26%) employed in the communicaion are multi-derived verbs. Concerning the second question. Is there a notable difference in this phenomenon between written and oral communication? We have found that verb forms containing two or more derivation suffixes is more frequent in an oral discourse(26%) than in a written one(8,8%)
We plan to continue our tests on larger corpora to confirm or not our hypothesis on written context.

References

Dubois J. and Dubois-Charlier F. (2007). Dictionnaire Linguistique et Sciences du langage, Editions Larousse, Paris.
Dubois J and Dubois-Charlier F. (1997). *Les verbes français*, Larousse, Paris.
Duran M. (2013). Formalizing Quechua verbs Inflexion. Proceedings of the NooJ 2013 International Conference, Saarbrücken.
Duran M. (2014). *Morphological and Syntactic Grammars for the Recognition of Verbal Lemmas in Quechua.* Proceedings of the NooJ 2014 International Conference, Sassari.
Duran M. (2017). Dictionnaire électronique français-quechua des verbes pour le TAL. Université Franche-Comté, Besançon-Paris. Thèse doctoral soutenu en mars, 600 p.
Itier C. (2003). Transcription des émissions de Radio Quispillaqta : Llaqtanchikpa sunqun, 18-03-2003.

[10] See Duran 2017.

Silberztein M. (2010). La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. Langages 3/2010 (n° 179-180), pp. 221-241.
Silberztein M. (2016). Language formalization : the NooJ's Approach. Wiley Eds. London.

Improving the performance of word frequencies in authorship attribution

Maciej Eder

Institute of Polish Language – maciej.eder@ijp.pan.pl

Abstract
In this paper, I introduce a very simple method of computing relative word frequencies for authorship attribution and similar stylometric tasks. Rather than computing relative frequencies as the number of occurrences of a given word divided by the total number of tokens in a text, I argue that a more efficient normalization factor is the total number of *relevant* tokens only. The notion of relevant words includes synonyms and, usually, a few dozen other words in some ways semantically similar to a word in question. To determine such a semantic background, one of word embedding models can be used. The proposed method outperforms classical most-frequent-word approaches by a few percentage points.

Keywords: authorship attribution, stylometry, relative word frequencies, word vectors, semantic neighbors

1. Introduction

In a vast majority of stylometric studies, relative frequencies of the most frequent words (MFWs) are used as the language features to betray the authorial "fingerprint". A vector of such relative word frequencies is then passed to one of the multidimensional machine-learning classification techniques, ranging from simple distance-based lazy learners, such as Delta (Burrows, 2002; Evert *et al*., 2017), to sophisticated neural network setups (Gómez-Adorno *et al*., 2018).
Even if alternative types of features have been introduced to authorship attribution (Peng *et al*., 2003; Hirst and Feiguina, 2007; Lučić and Blake, 2013) and tested in controlled experiments (Eder, 2011), the standard approach relying of word frequencies continues to be predominant in the field (Grieve, 2007; Stamatatos, 2009). In this paper, I propose to count the relative frequencies in a slightly different way, in order to better capture the authorial choice of words.

2. Words that (might) matter

The notion of relative word frequencies is fairly simple. We count all the tokens belonging to particular types (e.g. all the attestations of the word "the", followed by the attestations of "in", "for", "of" etc.), and for each word, we divide the number of types by the total number of words in a document. Consequently, each word frequency is equal to its percentage within the document (e.g. "the" = 0.0382), and all the frequencies sum up to 1. The reason of converting occurrences to relative frequencies is obvious: by doing so, one is able to reliably compare texts that differ in length. The notion of relative word frequencies is so natural and intuitive that one might very easily overlook its methodological implications, as if it was nothing else than a simple normalization procedure.

For the sake of this paper, it is important to realize that relative frequencies are relative to *all the other words* in a document in question. Convenient as they are, these values are at the same time very small and – importantly – they are affected by hundreds of other word frequencies. Consequently, the final values might not be sufficiently precise to capture minute differences between word occurrences, because the normalization factor evens them out to some extent. Now, what if we disregard thousands of other words in a text, and instead compute the frequencies in relation to a small number of words that are relevant? An obvious example is the mutual relation between the words "on" and "upon" in one document (Mosteller and Wallace, 1964); essentially, more attestations of "upon" come at the cost of the occurrences of the word "on" – and vice versa. While the classical relative frequency of the word "on" in Emily Bronte's *Wuthering Heights* is 0.00687, the proportion of "on" relative exclusively to "upon" is 0.9762. It is assumed in this paper that the latter frequency can betray the authorial signal to a greater extent than the classical approach, because the myriads of other words are not blurring the final value.

Given the above assumption, it would be tempting to identify one synonym for each of the words, and to compute the relative proportions in each of the synonym pairs (Borski and Kokowski, 2021). Linguistically speaking, however, such an approach would hardly be feasible. Firstly, only a fraction of words have their proper synonyms. Secondly, some semantic fields are rather rich and cannot be reduced to a mere pair of synonyms. Thirdly, in the case of the most frequent words (articles, particles, prepositions) identifying their synonyms doesn't make much sense, yet still, *relevant counterparts* for these frequent words obviously exist. On theoretical grounds, however, it is difficult to speculate whether the number of relevant counterparts should be restricted to a single word – as in the example of "on" defined by its relation to "upon" – or include, say, a dozen related words. E.g., to determine the relative frequency of the word "make", one could measure its proportion against the sum of occurrences of "do", "prepare", "create", "turn", "craft", "invent" etc. For this reason, an experimental

framework was set up in order to systematically assess tighter (1 relevant counterpart) and broader semantic background (up to 10,000 words, inevitably going far beyond synonyms). Rather than identifying rigid lexical synonyms, I used a word embedding model (GloVe, 100 dimensions) to extract *n* semantic nearest neighbors for each of the words in question. Consequently, the neighbors for the word "person" were: "woman", "gentleman", "man", "one", "sort", "whom", "thing", "young", etc., whereas the neighbors for the word "the" were as follows: "of", "this", "in", "there", "on", "one", "which", "its", "was", "a", "and", etc. For each target word, a relative frequency was calculated as the number of occurrences divided by the sum of occurrences of its n semantic neighbors (*n* being the size of semantic space to be tested).

3. Results

In order to corroborate the above intuitions, a controlled experiment was designed. A benchmark corpus of 100 English novels was used: it consists of 33 authorial classes and 3 novels per author, and is freely available on GitHub repository: https://github.com/computa-tionalstylistics/100_english_novels. The tests were performed using the package stylo for R (Eder *et al*., 2016). Different classifiers, MFW vectors and, most importantly, different sizes of the semantic space were tested systematically, in a supervised setup with stratified cross-validation. On theoretical grounds, the size of the semantic space $n = 37{,}000$ (roughly the total number word types in the benchmark corpus) would be equivalent to classical relative frequencies, whereas the space of the size $n = 1$ means that the frequencies are relative to exactly one other word (e.g. the frequency of the word "the" would be the number of occurrences of "the" divided by the total number of "the" and "of").

The obtained results clearly suggest that the new method outperforms the classical relative frequencies solution, no matter which distance measure is used. In agreement with several previous studies, longer MFW vectors worked better than, say, 100 MFWs. Also according to expectation, Cosine Delta proves to be the winner among the classifiers. Counter-intuitive was the behavior of different classifiers with the enhanced word frequencies. As evidenced in Fig. 1, Cosine Delta works best with frequencies computed against 5–50 semantically similar words, whereas Burrows's Delta (Fig. 2) exhibits its sweet spot for 50–100 neighboring words.

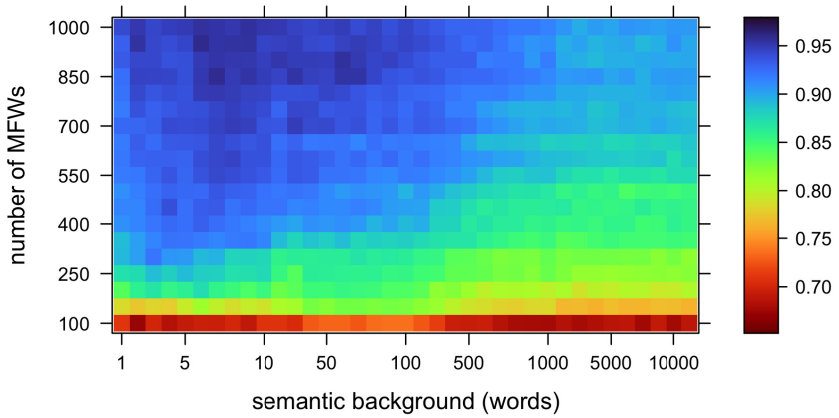


Figure 1. The performance (F1 scores) for a benchmark corpus of English novels and Cosine Delta as a classifier. The results depend on the MFW vector (y axis) and the size of the semantic space (x axis).

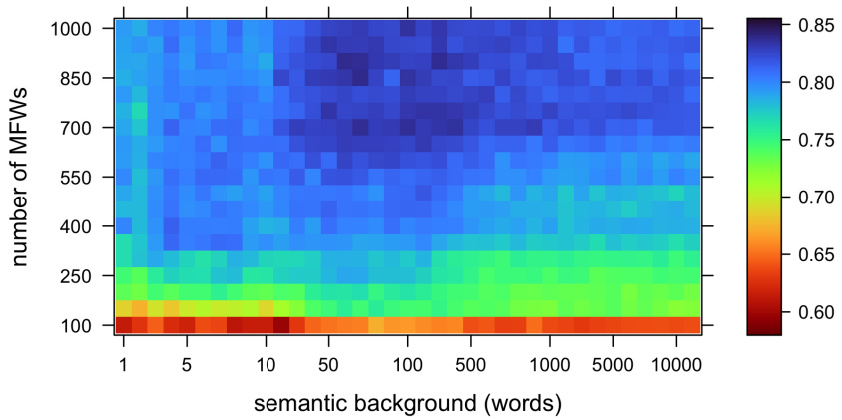


Figure 2. The performance (F1 scores) for a benchmark corpus of English novels and Burrows's Delta as a classifier.

According to the above results, a recipe for a successful authorship attribution setup seems to be as follows: take roughly 800–900 MFWs, and compute their frequencies using, for each word, the occurrences of their 5–10 semantic neighbors; then use the Cosine Delta classifier.

Since in authorship attribution the results are proven to be unevenly distributed across different MFW vectors, Fig. 3 presents the performance of the Cosine Delta model again, this time as the improvement (in percentage points) over the baseline. While the overall best performance is obtained for *ca*. 850 MFWs computed against 5–10 words, the biggest gain over the baseline (more than 10 percentage points!) is provided by the following scenario: 300 MFW frequencies

computed against a tight semantic background of 3 neighboring words. Other reasonable improvements are generally associated with short MFW vectors and the semantic background of 5–100 words. In the case of Burrows's Delta, which worked with 900 MFWs computed against 60 neighboring words (Fig. 2), the improvement over the baseline (not shown) is biggest for short vectors of MFWs. Interestingly, for Burrows's Delta the new method proves to be *worse* than the baseline for long MFW vectors and tight semantic spaces of 1–10 neighboring words. (The results which not shown here, are available at the project's GitHub repository).

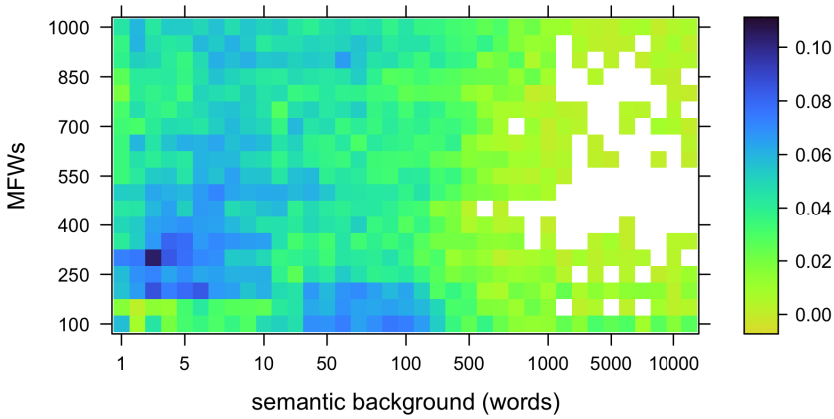


Figure 3. The gain (for Cosine Delta) over the classical relative frequencies, for different MFW vectors (y axis) and the size of the semantic space (x axis).

The proposed way of identifying an arbitrarily chosen number of semantic neighbors, might suffer from an uneven distribution of semantic neighbors in a given model (GloVe, word2vec, fastText, etc.). E.g., 50 neighboring lexemes might point to a semantically coherent area around a function word, or indicate but vague associations around a very specific technical term. To account for this factor, a second experiment has been conducted, in which I have defined a semantic background to be all the words located at a specific cosine distance from a given reference word. Consequently, rather than extracting *n* neighboring words, now I was extracting all the words within the radius of 0.9 cosine similarity in the first iteration, then 0.85, 0.8 etc. all the way to –0.9. The results for the Cosine Delta classifier are shown in Fig. 4. As can be seen, a clear hot spot forms in the area of 0.7–0.5 cosine similarity, and after the distance of 0.3 the performance hits the baseline level. The results confirm the general picture obtained in the previous experiment (Fig. 1), yet the sweet spot area seems to be more difficult to generalize.

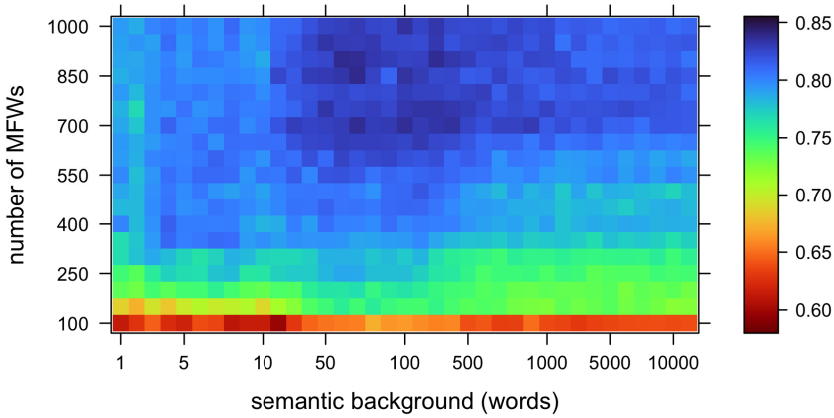


Figure 4. The performance (F1 scores) for a benchmark corpus of English novels and Cosine Delta as a classifier. The semantic space defined as the words within a given cosine distance from respective source words. The results depend on the MFW vector (y axis) and the size of the semantic space (x axis).

4. Conclusion

The paper presented a simple method to improve the performance in different stylometric setups. The method is conceptually straightforward and does not require any NLP tooling. The only external piece of information that is required is a list of semantically related words for each of the most frequent words in the corpus. A controlled experiment showed a significant improvement of classification accuracy in a supervised multi-class authorship attribution setup. A few general remarks can be formulated:

(1) No matter which classification method was used, the performance improvement turned out to be large, clearly suggesting that bare word occurrences retain much more authorial signal than the time-proven relative frequencies are able to betray. It can be safely hypothesized that the method introduced in this paper barely opened a new perspective, rather than offering an ultimate solution to the problem.

(2) In order to identify *the words that matter*, a word embedding model was used – and this, again, was far from an optimal solution. As a rough proxy, it nevertheless was able to improve the word frequencies in the range of 5–50 neighboring words. On theoretical grounds, a further improvement should be possible with a more precise method of identifying relevant semantic background.

(3) While the new method improves the performance across all the MFW strata, short MFW vectors seem to benefit more. Interesting from a theoretical point of view, this phenomenon has also a practical implication. Namely, since several

studies suggest that larger numbers of MFWs should be preferred as they generally exhibit better performance, it is also believed that they are more likely to be affected by genre, topic, and content of the analyzed texts. With this in mind, some practitioners choose to conduct authorship attribution on shorter MFW vectors. The method introduced in this paper can greatly improve the performance in such setups.

An observation that requires further investigation, is the discrepancy between classifiers in the way in which they react to the same semantic background. Contrary to intuition, for Burrows's Delta the improvement of performance was not simply correlated with the size of the semantic background. Tight neighborhood – less than 20 synonyms and/or other related words – did not outperform standard relative frequencies, whereas broader contextual information of ca. 50–100 related words showed a significant improvement over the baseline. In the case of Cosine Delta, tight semantic background of ca. 5–10 proved optimal, whereas broader spaces of 50–100 neighboring words were only marginally worse, still outperforming the baseline to a significant degree.

Acknowledgements

This research is part of the project 2017/26/E/HS2/01019, supported by Poland's National Science Centre. The code and the datasets to replicate the experiments presented in this study are posted on GitHub repository: https://github.com/computationalstylistics/word_frequencies.

References

Borski G. and Kokowski M. (2021). Copernicus, his Latin style and comments to Commentariolus. *Studia Historiae Scientiarum*, 20: 339–438.

Burrows J. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.

Eder M. (2011). Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6: 99–114.

Eder M., Rybicki J. and Kestemont M. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, 8(1): 107–21.

Evert S., Proisl T., Jannidis F., Reger I., Pielström S., Schöch C. and Vitt T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl. 2): 4–16, doi:10.1093/llc/fqx023.

Gómez-Adorno H., Posadas-Durán J.-P., Sidorov G. and Pinto D. (2018). Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7): 741–56, doi:10.1007/s00607-018-0587-8.

Grieve, J. W. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary And Linguistic Computing*, 22(3): 251–70, doi:10.1093/llc/fqm020.

Hirst G. and Feiguina O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4): 405–17.

Lučić A. and Blake C. L. (2013). A syntactic characterization of authorship style surrounding proper names. *Digital Scholarship in the Humanities*, 30(1): 53, doi:10.1093/llc/fqt033.
Mosteller F. and Wallace D. (1964). I*nference and Disputed Authorship: The Federalist*. Stanford: CSLI Publications.
Peng F., Schuurmans D., Keselj V. and Wang S. (2003). Language independent authorship attribution using character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 267–74.
Stamatatos E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538–56.

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

Louis Escouflaire, Antonin Descampe, Cédrick Fairon

UCLouvain – [firstname.lastname]@uclouvain.be

Abstract
The presence of subjectivity in news articles has become a more salient issue since the advent of social networks as platforms for the distribution of news (Tucker *et al*., 2018). To examine if subjectivity has evolved in French-speaking journalism in recent years, we analyze 120.000 articles published between 2010 and 2021 by the RTBF, the Belgian public service media, using a score based on several linguistic measures of subjectivity.

Keywords: Subjectivity analysis, journalism studies, corpus linguistics, NLP.

Résumé
La présence de la subjectivité dans les articles d'information est devenue une problématique plus saillante depuis l'avènement des plateformes socio-numériques en tant que lieux de diffusion de l'information (Tucker *et al*., 2018). Afin d'examiner si la subjectivité a évolué dans le journalisme francophone ces dernières années, nous analysons 120 000 articles publiés entre 2010 et 2021 par la RTBF, le média de service public belge, à l'aide d'un algorithme utilisant plusieurs mesures linguistiques de la subjectivité.

Mots clés : Analyse de subjectivité, journalisme, linguistique de corpus, TAL.

1. Introduction

D'après une étude récente (Matsa *et al*., 2018), un jeune Européen sur deux s'informe tous les jours via les plateformes socio-numériques comme Facebook ou Instagram. La même étude témoigne d'une perte de confiance envers les médias traditionnels et de service public. Alors que les algorithmes de recommandation utilisés pour diffuser l'information sur ces plateformes favorisent la mise en avant d'articles suscitant l'engagement émotionnel des utilisateurs, articles plus enclins à être empreints de subjectivité (Tucker *et al*., 2018), il devient intéressant d'observer comment les médias traditionnels se sont adaptés à ce nouveau paradigme.

Dans cet article, nous analysons l'évolution de la subjectivité dans les productions journalistiques écrites de la Radio-Télévision Belge Francophone (RTBF), un média de service public publiant chaque jour des centaines d'articles de presse sur son site web. A l'aide d'un score de subjectivité constitué d'un assemblage de mesures linguistiques issues de la littérature et à partir d'un corpus de 120 000 articles publiés entre 2010 et 2021, période cruciale traversée par l'explosion des plateformes socio-numériques, nous chercherons à examiner si le caractère subjectif des articles d'information de la RTBF a évolué sur douze années. Notre hypothèse initiale est que les articles web pourraient être devenus plus subjectifs au cours du temps, afin de gagner en visibilité sur ces plateformes qui tendent à favoriser les articles les plus polarisants, qui sont aussi souvent les plus subjectifs (Tucker *et al*, 2018).

2. État de l'art

Bien que l'objectivité soit souvent considérée par les journalistes comme une norme pour l'écriture d'articles d'information, il est communément admis qu'elle est un idéal inatteignable (Steensen, 2017). Plusieurs étapes essentielles du travail de journaliste, comme la sélection des informations, la hiérarchisation des sources, ou simplement la description des faits, sont des tâches indissociables des choix personnels faits par les auteurs, irrémédiablement influencés par leurs opinions personnelles et par leurs expériences passées (Koren, 2004). Ce n'est qu'en se soumettant à ce que Tuchman (1972) appelle le « rituel de l'objectivité », donc en recourant à des procédés linguistiques comme l'emploi de mots « neutres » ou de tournures impersonnelles, que les journalistes peuvent masquer l'impact de leur subjectivité sur leur texte (Charaudeau, 2006). En linguistique, la subjectivité est considérée par beaucoup de linguistes comme une propriété inhérente au langage, notamment depuis les travaux de Benveniste (1966). Dans le discours, elle prend la forme de marqueurs linguistiques qui traduisent l'opinion ou l'évaluation de l'énonciateur sur les réalités dont traitent le discours, ou sur le discours lui-même (Hunston & Thompson, 2000). Catherine Kerbrat-Orecchioni (1980) a recensé les différents types de traces énonciatives, ou « unités subjectives », qui peuvent apparaître en français. Plus tard, dans le champ du traitement automatique du langage (TAL), les travaux de Wiebe *et al*. (2004) étendent cette vision composite de la subjectivité dans le discours en mesurant le caractère subjectif d'un texte ou d'une phrase à partir du nombre d'unités subjectives qu'il contient. Parmi les marqueurs linguistiques de la subjectivité identifiés comme pertinents par différentes études, on trouve des mesures liées à la distribution des catégories morphosyntaxiques dans le texte, des mesures lexicosémantiques basées sur des lexiques de mots subjectifs, ou des mesures stylométriques concernant par exemple la ponctuation, la longueur des mots ou le nombre de citations (Krüger *et al*., 2017).

L'analyse de la subjectivité dans les articles de presse a été effectuée de diverses manières, en anglais comme en français (Ho-Dac & Küppers, 2011 ; Krüger *et al*., 2017 ; Todirascu, 2019). D'un point de vue longitudinal, Charron (2006) a constaté que, de 1960 à 1970, les articles de presse en français du Québec avaient significativement gagné en subjectivité. Chaput (2019), en se basant uniquement sur des mesures stylométriques et sur un corpus de 1.200 articles, a montré que la présence de l'énonciateur dans le journalisme québécois se faisait de plus en plus grande entre 1945 et 2015. Notre étude cherche à examiner l'évolution de la subjectivité dans les articles web du service public belge francophone entre 2010 et 2021, et se distingue notamment des études précédentes par l'ampleur du corpus utilisé (120 000 articles) et par la variété des indicateurs combinés pour analyser la subjectivité.

3. Méthodologie

3.1. Corpus

Les articles utilisés sont issus des archives de la RTBF, contenant à ce jour plus de deux millions d'articles de presse publiés quotidiennement par le média de service public belge francophone depuis 2008. Nous analysons un échantillon composé d'articles du canal *information* publiés entre 2010 et 2021, issus des catégories *monde*, *Belgique*, *régions*, *société* et *économie*. Cette présélection nous permet de travailler sur un ensemble d'articles partageant une certaine homogénéité thématique. Constatant que le nombre moyen d'articles publiés chaque année par la RTBF n'est pas constant, comme nous le verrons plus loin, nous avons extrait au hasard 10.000 articles par année, de 2010 à 2021, constituant un corpus total de 120 000 articles.[1]

3.2. Score de subjectivité

Afin de situer chacun des 120 000 articles sur une échelle graduelle allant de l'article le plus « factuel » (score le plus bas) à l'article le plus « subjectif » (score le plus haut), un score de subjectivité a été conçu puis normalisé afin d'attribuer à chaque article un score entre 0 et 1.

Ce score de subjectivité est élaboré sur base des résultats d'une étude préalable (Escouflaire, 2022) au cours de laquelle 30 indicateurs linguistiques de la subjectivité tirés de l'état de l'art ont été évalués au travers d'une tâche de classification de 13 400 articles étiquetés par la RTBF comme des articles d'information et d'opinion de la RTBF. Au cours de cette expérience, 18

[1] Des discussions sont en cours pour rendre le corpus disponible à des fins de reproductibilité des résultats.

indicateurs ont été identifiés comme efficaces pour la classification, par un modèle de régression logistique binomiale. A partir des 18 mesures linguistiques, nous avons mis au point ce score de subjectivité, qui est une somme de ces mesures calculées sur base du texte de l'article, pondérée en fonction du coefficient de régression accordé à chaque mesure par le modèle de classification. Ce modèle et les pondérations exactes attribuée à chaque indicateur sont présentés en détails dans l'article consacré à la sélection des indicateurs (Escouflaire, 2022). Le Tableau 1 énumère les 18 indicateurs qui sont mesurés et combinés pour obtenir le score : à gauche, les indicateurs en faveur des articles plus subjectifs, à droite les indicateurs qui tendent à rendre l'article moins objectif. Chaque mesure est normalisée en fonction de la longueur totale de l'article.

Tableau 1. Les 18 indicateurs combinés pour construire le score de subjectivité.

+ subjectif		**- subjectif**
Nb de *on* (pronom indéfini)	Nb de pronoms relatifs	Nb de chiffres
Nb de mots de négation	Nb d'adjectifs	Nb de pronoms (1e p.)
Nb de mots d'émotion (*Lex3*)[2]	Nb de mots de sentiment (*NRC*)[2]	Nb de verbes
Nb de points d'exclamation (!)	Nb de points d'interrogation (?)	Longueur moy. des mots
Nb de points-virgules (;)	Nb de points de suspension (…)	Nb de citations
Nb de deux-points (:)	Taux de sentiment (*TextBlob*)[2]	
CTTR[2]		

4. Résultats et discussion

Une fois notre score de subjectivité appliqué sur les 120 000 articles d'information du corpus, il apparait que 95% des articles ont un score de subjectivité entre 0.62 et 0.79, et que 50% ont un score entre 0.68 et 0.74 (voir Figure 1). Après exploration des résultats, nous constatons que les articles considérés les moins subjectifs sont des comptes rendus financiers ou des dépêches très brèves, tandis que les articles avec le score de subjectivité le plus élevé sont des interviews ou des articles d'opinion qui n'ont pas été correctement étiquetés par la RTBF, ce qui consolide la puissance prédictive de notre score. Ces articles dont le score fait exception seront examinés de manière qualitative et avec plus de précision dans des travaux futurs. Par exemple, un article publié en 2017, décrivant la campagne médiatique de Donald Trump[3], comporte un nombre conséquent de traits linguistiques de la subjectivité, utilisés dans notre score : emploi du on, présence

[2] Les mesures de sentiment lexicosémantiques utilisées sont le lexique de valence émotionnelle de Lexique3 (Gobin *et al*., 2017), la traduction du lexique NRC EmoLex (Mohammad & Turney, 2013) et le module Python *TextBlob* (Loria, 2018). Le CTTR est une version corrigée du *type-token ratio*.

[3] https://www.rtbf.be/article/stop-aux-fake-news-donald-trump-diffuse-the-real-news-presentees-par-sa-belle-fille-9675818 (consulté le 16 mai 2022).

de mots issus des lexiques de sentiment (*défavorable, fiasco, accuse*) et utilisation de points de suspension et d'exclamation.

Ensuite, en examinant la courbe présentée dans la Figure 2, nous observons que le score de subjectivité moyen semble augmenter de manière constante entre 2010 et 2013, avant de se stabiliser jusqu'en 2017, il descend fortement entre 2017 et 2018. Cette baisse du score se poursuit, bien que moins nettement, jusqu'en 2021. Pour tenter d'expliquer ces résultats inattendus car contraires à notre hypothèse initiale, nous vérifions d'abord si le score de subjectivité d'un article est corrélé ou non avec sa longueur (mesurée en nombre de caractères). En effet, la taille moyenne des articles a augmenté de manière presque ininterrompue entre 2010 et 2018 (passant de 1600 à plus de 2000 caractères en moyenne). Un test de Pearson montre que la longueur d'un article n'influence pas significativement le score de subjectivité qui lui est attribué (r = -0.53, p = 0.09). Une autre explication de cette évolution de la subjectivité pourrait tenir dans le fait que les articles du corpus proviennent de sources différentes. Après

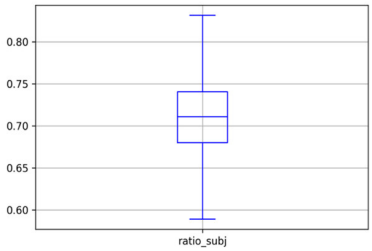


Fig. 1. Distribution du score de subjectivité de 95% des 120.000 articles d'information.

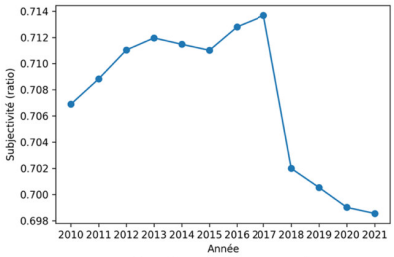


Fig. 2. Score de subjectivité moyen des 120.000 articles entre 2010 et 2021.

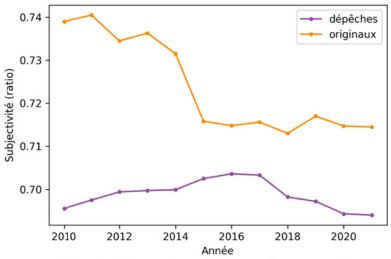


Fig. 3. Score de subjectivité moyen des dépêches et originaux entre 2010 et 2021.

investigation, nous constatons que 62.7% des articles présents dans le corpus sont des dépêches issues des agences de presse Belga et AFP (publiées telles quelles ou légèrement retravaillées par les journalistes RTBF), et que leur nombre varie fortement d'année en année. Avant 2015, environ 70% des articles publiés par la RTBF sont des dépêches. Entre 2015 et 2017, celles-ci ne représentent plus que 10 à 35% des articles. Puis, à partir de 2018, plus de 80% des articles publiés sont des dépêches. Notre intuition est que les dépêches sont par nature

moins subjectives que les articles « originaux » (écrits et publiés par des journalistes RTBF). Cette hypothèse est confirmée par un test de Student (t = 73.86, $p < 0.001$) : les articles originaux possèdent un score de subjectivité significativement plus élevé que les dépêches relayées par le média. Afin de décortiquer nos résultats, nous examinons séparément l'évolution de la subjectivité dans les dépêches et dans les articles originaux (Figure 3).

Dans la Figure 3, nous observons que le score de subjectivité moyen des dépêches est plutôt stable, puisqu'il n'oscille que sur 1% en douze ans. Pour les articles originaux, écrits intégralement par des journalistes RTBF, une baisse de la subjectivité se manifeste entre 2010 et 2014, s'accélérant nettement entre 2014 et 2015. En moyenne, les articles post-2015 sont significativement moins subjectifs que les articles pré-2015 (H = 72561.53, $p < 0.001$).

Deux entretiens oraux avec la RTBF, le premier avec le directeur adjoint de l'information actuel, le second avec le chef de la rédaction web de 2008 à 2017, nous permettent d'expliquer les différentes évolutions dans les Figures 3 et 4. A partir de fin 2013, un changement de paradigme rédactionnel s'est opéré au sein du média de service public belge : le passage au mode « bimédia ». Avant 2013, une équipe d'une dizaine de rédacteurs spécialisés dans l'écriture web se chargeait d'écrire tous les articles pour le site web. Afin d'augmenter le nombre d'articles publiés sur leur site Internet, la RTBF a élargi la tâche de rédiger des articles web à un grand nombre de ses journalistes TV et radio, devenus des journalistes bimédias. La transition vers cette logique rédactionnelle « à 360° » implique une explosion du nombre d'articles originaux de la RTBF, ce qui explique la baisse du nombre proportionnel de dépêches entre 2013 et 2015. Nos entretiens ont révélé qu'au même moment, la RTBF a fait diffuser auprès de ces nouveaux rédacteurs, non spécialisés dans la production de contenu web, un document contenant un ensemble de lignes directrices à appliquer pour la rédaction d'articles.

Les recommandations énoncées dans ce document, que nous avons pu consulter, traduisent deux volontés de la part du média. D'une part, il est demandé aux journalistes d'adapter autant que possible leur écriture aux normes éditoriales de la RTBF : éviter la ponctuation expressive, limiter le nombre de négations, réduire l'emploi du pronom *on*, entre autres. Certaines de ces consignes reflètent exactement certains des indicateurs utilisés dans notre score de subjectivité, présentés dans le Tableau 1 : moins un article contient de points d'interrogation, d'exclamation, de mots de négation ou d'occurrences de on, moins il est considéré comme subjectif par notre score. Ces éléments expliquent donc la baisse drastique du score de subjectivité à partir de 2014 dans les articles originaux. D'autre part, le document attend des rédacteurs qu'ils appliquent plusieurs stratégies rédactionnelles censées améliorer le référencement des articles dans les moteurs de recherche ou les algorithmes de recommandation : en publiant des articles plus longs (ce qui explique nos premières observations),

en favorisant le contenu original (mieux référencé), ou en enrichissant les dépêches avec des ajouts originaux. Le pic de subjectivité observé en 2014 dans les dépêches, ainsi que la baisse proportionnelle du nombre de dépêches après cette date, peuvent être expliqués par l'application de ces conseils : étant plus systématiquement retravaillées par les journalistes RTBF, les dépêches deviennent plus susceptibles de contenir de la subjectivité.

Un second phénomène rapporté au cours de nos entretiens peut expliquer la courbe surprenante de la Figure 2. En 2017, le chef de la rédaction web de la RTBF a été remplacé, ainsi qu'une partie de son équipe de rédacteurs. La politique bimédia a été maintenue, expliquant la constance dans la subjectivité des articles originaux jusqu'à aujourd'hui, mais le mode de publication des dépêches a été remodelé : la RTBF a recommencé à relayer sur son site web un grand nombre de dépêches peu ou pas du tout modifiées, leur publication étant parfois automatisée (pendant la nuit, par exemple). Entre 2017 et 2018, le nombre total d'articles d'information publiés sur le site de la RTBF passe d'environ 30 000 à 120 000 par an. La hausse du nombre de dépêches publiées de manière plus brute, sous une forme a priori moins subjective, explique la baisse du score de subjectivité des dépêches observée dans la Figure 3, ainsi que sa baisse globale dans les articles RTBF (Figure 2), les dépêches représentant plus de 80% des articles publiés entre 2018 et 2021.

5. Conclusion

Nos résultats ont dévoilé une baisse significative de la subjectivité dans les articles d'information de la RTBF depuis 2015, en particulier dans les articles non issus de dépêches. Ces observations semblent aller dans le sens de notre hypothèse initiale selon laquelles les caractéristiques fonctionnelles des algorithmes de recommandation sur le web ont influencé les pratiques d'écriture journalistique, mais sont surprenantes puisque nous nous attendions plutôt à une montée de la subjectivité au cours des dernières années. Grâce à nos entretiens avec deux responsables éditoriaux de la RTBF, nous avons pu comprendre que les résultats obtenus à l'aide de notre mesure de subjectivité reflètent des mouvements rédactionnels internes visant à améliorer la visibilité en ligne des articles du média belge. D'autre part, cette évolution peut aussi être vue comme une homogénéisation de l'écriture journalistique, sur base de critères imposés indirectement par les algorithmes de référencement.

Ces résultats ouvrent plusieurs perspectives intéressantes. D'abord, pour poursuivre nos analyses, il serait pertinent d'étudier plus précisément lesquels des 18 indicateurs linguistiques ont le plus d'influence sur l'évolution attestée du score de subjectivité. Ensuite, en reproduisant nos analyses sur un corpus d'articles d'autres médias francophones, nous souhaitons évaluer si ces observations sont uniquement dues à des décisions internes à la RTBF ou si elles

peuvent être généralisées à d'autres espaces de diffusion que le web, ou à d'autres rédactions francophones. Enfin, dans le domaine des journalism studies, l'homogénéisation décrite ci-dessus étaye des pistes de recherche intéressantes relatives à la perte de souveraineté éditoriale subie par les médias d'information et influencée par les modes de diffusion des plateformes socio-numériques.

Bibliographie

Benveniste E. (1966). De la subjectivité dans le langage. *Problèmes de linguistique générale*, Paris, Gallimard (coll. Bibliothèque des sciences humaines), 258-266.

Chaput L. (2019). Sur quelques marques de subjectivité dans le journalisme d'information politique de 1945 à 2015 au Québec. Mots, 119, 151-168.

Charaudeau P. (2006). Discours journalistique et positionnements énonciatifs. Semen, 22.

Charron J. (2006). Subjectivation du discours du journalisme politique au Québec, 1945-1995. *Mots. Les langages du politique*, (82), 81-94.

Escouflaire L. (2022). Identification des indicateurs linguistiques de la subjectivité les plus efficaces pour la classification d'articles de presse en français. *TALN-RECITAL 2022*.

Gobin P., *et al*. (2017). Une base de l'émotionalité (valence, arousal, catégories) de 1286 mots français selon l'âge (EMA). *European Review of Applied Psychology*, 67(1), 25-42.

Ho-Dac L.-M. & Küppers A. (2011). La subjectivité à travers les médias : Étude comparée des médias participatifs et de la presse traditionnelle. *Corpus, 10*, 179-199.

Hunston S. & Thompson G. (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford University Press, UK.

Kerbrat-Orecchioni C. (2009). *L'énonciation : de la subjectivité dans le langage*. Armand Colin.

Wiebe J., *et al*. (2004). Learning Subjective Language. *Computational Linguistics*, 30(3), 277-308.

Koren R. (2004). Argumentation, enjeux et pratique de l'"engagement neutre" : Le cas de l'écriture de presse. Semen, 17.

Krüger K. R., *et al*. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5), 687-707.

Loria S. (2018). TextBlob Documentation. Release 0.15, 2.

Matsa, K. E., Silver, L., Shearer, E., & Walker, M. (2018). *Western Europeans Under 30 View News Media Less Positively, Rely More on Digital Platforms Than Older Adults*. Pew Research Center.

Mohammad S. & Turney P. (2013). Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence*, 29 (3), 436-465.

Steensen S. (2017). Subjectivity as a Journalistic Ideal. Putting a Face on it: Individual Exposure and Subjectivity in Journalism. *Cappelen Damm Akademisk*, 25-47.

Todirascu A. (2019). Genre et classification automatique en TAL : Les genres journalistiques. *Linx, 78*.

Tuchman G. (1972). Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology*, 77(4), 660-679.

Tucker J. A., *et al*. (2018) *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature*. Rochester, NY: Social Science Research Network.

The economy and the Web: RETI a regional economic twitter index

Shira Fano[1], Gianluca Toschi[2]

[1]Fondazione Nord Est – shira.fano@fnordest.it
[2]Fondazione Nord Est and IRCRES-CNR – gianluca.toschi@fnordest.it

Abstract
In recent years, data from social media are increasingly used to monitor opinions, forecast, and analyse economic and social dynamics. In this paper we develop RETI, a Regional Economic Twitter Index capable of analysing economic dynamics at the regional level, monitoring daily posts on the social media Twitter. Starting from April 2020, we automatically downloaded tweets referred to the North-East of Italy and containing economic terms, selected using state of the art word-embedding techniques. We applied Sentiment Analysis techniques, using an unsupervised lexincon-based approach and computed the economic sentiment of users. The index dynamics is coherent with real world facts and media information and is positively correlated with standard economic indicators computed by ISTAT, confirming the ability of social media of capturing facts and short run confidence in the economy.

Keywords: Regional index; Twitter data; Word-embedding; Sentiment Analysis.

1. Introduction

In the study of regional economic dynamics, there are two main problems: the first is the lack of infra-annual data, the second is the delay with which regional statistics are released by ISTAT. Using data from the web and social media could, therefore, represent a possible solution. In recent years, data from social media are increasingly used to monitor opinions, forecast, and analyse economic and social dynamics, (Antenucci, D. *et al*., 2014). Combined with traditional data collection methods, such as surveys, social media help monitor phenomena effectively in real-time manner. In this work, we develop RETI (Regional Economic Twitter Index) an index capable of analysing economic dynamics at the regional level, monitoring daily posts on the social media Twitter. Our index is inspired by the Economic Social Mood Index proposed by ISTAT but, differently from it, we focus on the regional level and improve the sentiment analysis algorithm. (Gallo M. *et al*., 2019) developed "VEN-ICE" for the Veneto region; the index estimates the evolution of regional GDP in the medium-long term monthly. Our indicator instead in based on real-time citizens' opinions collected via Twitter, without the intention of predicting GDP.

2. Data

Starting from April 1st, 2020, we have been downloading all tweets containing 1) the name of one of the four regions of *Veneto*, *Friuli-Venezia Giulia*, *Trentino-Alto Adige* and *Emilia-Romagna* or one of the provinces that compose them and 2) a term related to the economy selected from a pre-determined batch. We automatically downloaded tweets using the Twitter API through the R statistical programming language (https://www.r-project.org/) and we used the R package rtweet. This allows us to gather in a database form the text of the tweets, together with related information such as the exact time of the tweet, if it is a tweet or a retweet, and the device type from which the tweet is posted. Once the tweets have been collected, it is necessary to selected from the corpus only those related to economic sub-topics. To identify the economic topics, we created a set of 1400 words related to the economy and selected tweets containing at least one of these words. To identify economy-related lemmas, we applied a two-step procedure. First, we defined a list of more than 200 words related to the economy. To enrich this set, we then used *word-embedding* techniques, which allow words to be represented as vectors using various training methods inspired by neural-network language modelling. We use the training method developed by Mikolov *et al*. (2013) as implemented in the R package *word2vec*. Before analysing the text of tweets, we clean and pre-process the data, using the R packages for text mining tm and tidytext. We perform several steps that include creating a corpus, removing stop words, removing URLs, punctuation, and symbols, converting all words to lower case and finally stemming words to retrieve their root.

3. Methods

We analyse the text of economic tweets related to the economy in the Italian North-East and perform a Sentiment Analysis (SA) to gather information on the mood of Twitter users. There are three main methods used in recent literature to carry out sentiment analysis: machine learning, rule-based approach, and lexicon-based approach. Machine learning strategies are based on the training of an algorithm with a training data set. The rule-based approach defines various rules for capturing opinions, created by tokenising each sentence and every document. Lexicon-based techniques start from the assumption that the polarity of a sentence is the sum of polarities of the individual words (Devika *et al*., 2016). In our analysis, we use a lexicon-based approach based on Sentix, an Italian polarity lexicon. The advantage of using this sentiment polarity lexicon is that, compared to simpler lexicon-based approaches, it does not assign a binary value indicating if lemmas are either positive or negative, but it also indicates the degree of polarisation of lemmas. The sentiment analysis consists of two steps: first we compute a sentiment score for each tweet, second, we use the sentiment scores to cluster tweets into three mutually exclusive classes: positive (*P*),

negative (*N*) and neutral (*U*). To assign sentiment scores to a tweet we use an Italian Sentiment Lexicon, a vocabulary whose lemmas are associated with pre-computed *positive* and *negative* sentiment scores, see (Basile and Nassim, 2013). This allows assigning a positive or negative polarity to terms. The Italian polarity lexicon, called Sentix (SENTiment Italian lexicon) is uses. An entry in Sentix consists in an Italian lemma (noun, verb, adjective, adverb), a positive and a negative score ranging from 0 to 1, a polarity score ranging from -1 to 1 and an intensity score ranging from 0 to 1.
To allow clustering, also tweets must be mapped into a 4-dimensional space therefore given positive, negative, polarity and intensity values of matched words in each tweet we compute the average of positive and negative values and, from these, polarity, and intensity, allowing us to associate to each tweet these four values. Once sentiment scores (*p, n, ω, i*) are available for all the tweets of a daily block we use *K*-means algorithm to cluster them into Positive, Negative and Neutral tweets. Finally, we compute (S) the daily index value:

$$S = \overline{\omega}_i = \frac{\sum_t i_t\, \omega_t}{\sum_t i_t} = \frac{\sum_{t \in P} i_t\, \omega_t + \sum_{t \in N} i_t\, \omega_t}{\sum_t i_t} \qquad \text{where} \quad \omega_t \stackrel{\text{def}}{=} 0 \;\; \forall t \in \text{Neutral} \tag{1}$$

This index is the average of polarity (*ω*) weighted by intensity (*I*), provided we assign polarity 0 to neutral tweets. After the cluster analysis, the value of S, the sentiment for a given day, is computed. The procedure is repeated for all daily blocks, allowing us to obtain a time series for the economic sentiment.

4. Results and discussion

Table 1 is an example that helps to better understand how polarity of a given tweet is computed, from polarity and intensity of single lemmas. Consider for example the tweet: «*Gravissimo infortunio presso centro servizi ravenna cgil cisl uil proclamano sciopero*» (Severe injury in the service center in Ravenna *cgil cisl* and *uil* declare a strike). Given positive and negative values of *infortunio*, *centro* and *sciopero* the average values are computed and from these polarity and intensity:

Polarity (ω) = $1 - 4tan^{-1}\left(\frac{0{,}304}{0{,}07}\right)/\pi$ = - 0,712 and Intensity(*i*) = $\sqrt{0{,}07^2 + 0{,}304^2} = 0{,}312$.

The tweet is therefore mapped into the vector (0,07; 0,304; -0,712; 0,321) and has a negative meaning.

Table 1. Tweet classification

	Positive	**Negative**	**Polarity**	**Intensity**
infortunio	0	0,708	-1	0,708
centro	0,21	0,08	0,53	0,26
sciopero	0	0,125	-1	0,125
Tweet	**0,07**	**0,304**	**-0,712**	**0,312**

The scatter plot in Figure 1 displays, as an example, the result of the cluster analysis for one of the studied days, June 30th, 2021. Each datapoint represents a tweet defined by a positive value, on the *x*-axis and a negative value, on the *y*-axis. Tweets are colour-coded according to the results of the cluster analysis, and they confirm the ability of the cluster analysis to group tweets with similar positive and negative values.

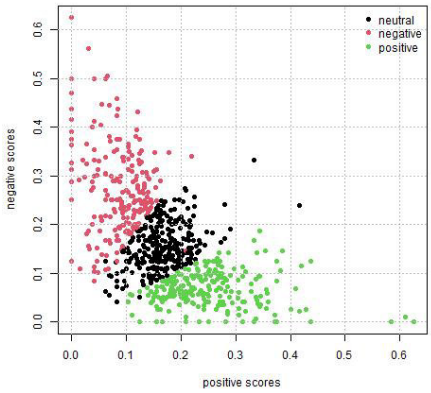


Figure 1. Daily clustering of tweets

The time series in Figure 2 describes the economic sentiment of twitter users about the Italian North-East in the studies period. The red line is a seven-day moving average that allows us to better visualize the index trend. Peaks and sudden drops correspond the real-world events confirming the ability of our indicator to nowcast economic activity and sentiment. One of the days with the lowest economic sentiment values is for example July 28th, 2020. On this day discussions on twitter focus on: *i*) car accidents and highway closed towards Trieste, *ii*) crisis in the house market in Emilia-Romagna and Rimini and *iii*) fake dollars disappearing from the police station in Piacenza. A positive day is for example June 20th, 2021, with new job offers in Treviso, in Verona and above

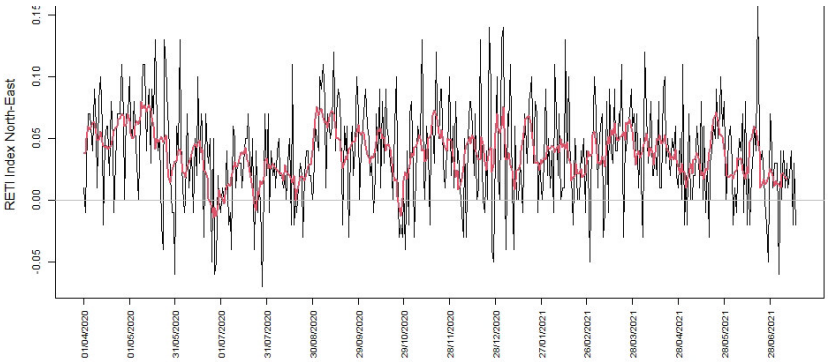


Figure 2. Regional Economic Twitter Index RETI for the Italian North-East

average employment rate in Parma.

Figure 3 displays the economic sentiment in the four single regions of *Emilia-Romagna, Veneto, Friuli-Venezia Giulia, and Trentino-Alto Adige* and, also in these cases peaks and drops area able to capture real world phenomenon and economic sentiment.

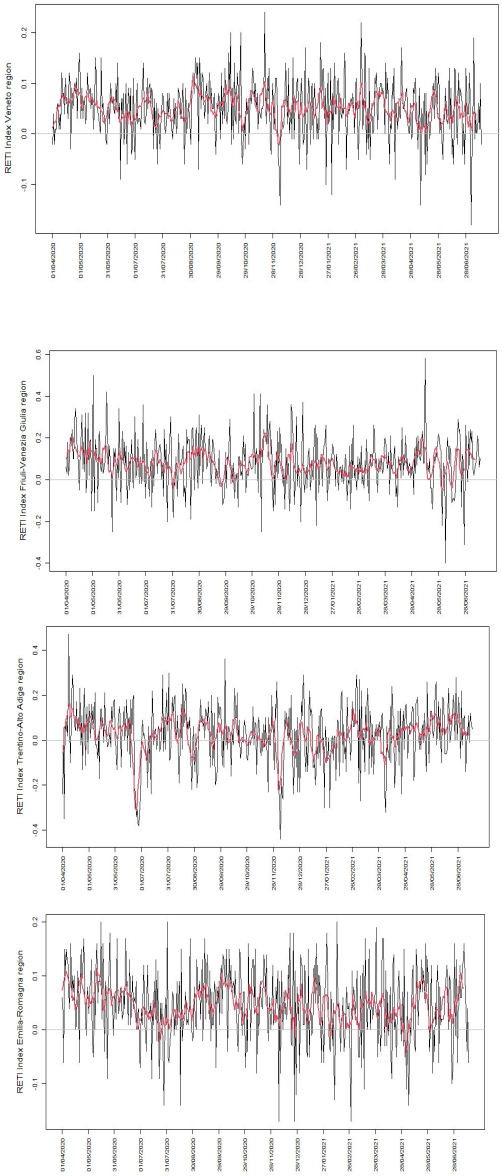


Figure 3. Regional Economic Twitter Index RETI for the Italian North-East

To test the validity of our indicator we considered the macro regional indicator RETI including the four regions and aggregated data at the monthly level. We then correlated the monthly macro-regional indicator RETI we obtained with the "Giudizio sulla situazione economica della famiglia" released by ISTAT on a monthly basis, based on a standard survey. The correlation, described in Figure 4, is positive and statistically significant (R^2 = 0,3), suggesting the validity of our non-standard indicator; this correlation should be further investigated in future research. Finally, the economic sentiment analysis successfully demonstrates the ability to capture news facts and user reactions, confirming the potential of this tool for regional economic studies.

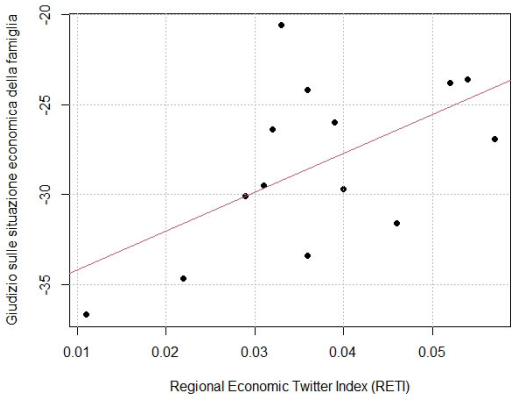


Figure 4. Monthly macro-region RETI and the "Opinion on the economic situation of families"

From a methodological point of view, the lexicon-based approach using Sentix allows to associate to lemmas and tweets a degree of polarization, not only binary values. Nevertheless, there some limitations that should be addressed in further research. We found the presence of some outliers that should be manually removed. Moreover, the lexicon-based approach has some limits, for example it cannot deal with irony, therefore supervised learning algorithms should also be considered.

References

Antenucci D., Cafarella M., Levenstein M., Ré C. and Shapiro M. D. (2014). Using social media to measure labor market flows (No. w20010). National Bureau of Economic Research.

Basile V. and Nissim M. (2013). Sentiment analysis on Italian tweets. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 100-107).

Devika M. D., Sunitha C. and Ganesh A. (2016). Sentiment analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44-49.
Freinerer I and Hornik K. (2020). A framework for text mining applications within R.
Kruspe A., Häberle M., Kuhn I., and Zhu X. X. (2020). Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. arXiv preprint arXiv:2008.12172.
Gallo M., Soncin S. and Venturini A. (2019). Ven-ICE: un nuovo indicatore delle condizioni dell'economia del Veneto (Ven-ICE: A New Indicator for the Economy of the Veneto Region). Bank of Italy Occasional Paper, (498).
Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Wiffels J. (2021). R package word2vec, Distributed Representation of words.

Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione

Manuel Favaro[1], Marco Biffi[2], Simonetta Montemagni[1]

[1]Istituto di Linguistica Computazionale "A. Zampolli" – CNR – manuel.favaro@ilc.cnr.it direttore@ilc.cnr.it
[2]Università di Firenze – Accademia della Crusca – marco.biffi@unifi.it

Abstract
This paper aims to show some of the resources developed within the TrAVaSI project (Automatic Treatment of Italian Historical Varieties). In particular, we illustrate corpora and morphological lexicons built to improve the performance of the lemmatization process, which still represents a challenge when dealing with historical varieties of language. Results achieved by extending the dictionary and by training the lemmatization model with historical corpora are compared. A preliminary analysis of lemmatization error typologies is also reported.

Keywords: NLP, corpora, morphological lexicons, historical varieties of language.

Riassunto
Il contributo si propone di illustrare alcune delle risorse sviluppate all'interno del progetto TrAVaSI (Trattamento Automatico di Varietà Storiche di Italiano). Ci si concentrerà sulla creazione di corpora annotati e lessici morfologici per il miglioramento dell'accuratezza del processo di lemmatizzazione. I risultati sono ottenuti attraverso l'estensione dei dizionari morfologici e l'addestramento del modulo di lemmatizzazione con corpora rappresentativi delle varietà storiche della lingua trattate. Sono presentati i risultati di uno studio preliminare sulle tipologie di errore di lemmatizzazione riscontrate nei diversi approcci.

Parole chiave: TAL, corpora, lessici morfologici, varietà storiche della lingua.

1. Introduzione

La disponibilità di archivi testuali digitali di rilevanza strategica per le discipline umanistiche cresce di giorno in giorno: al loro interno molti sono i testi che forniscono testimonianza di varietà storiche della lingua italiana. L'accesso, l'interrogazione e l'analisi dei testi all'interno di tali archivi rimangono spesso circoscritti alle stringhe di caratteri che costituiscono il testo. Nell'ultimo decennio, il settore delle Digital Humanities (DH) si è evoluto in modo

significativo, passando a sistemi avanzati di analisi ed elaborazione della lingua e del testo. Al loro interno, un ruolo centrale è rivestito da metodi e tecniche per il Trattamento Automatico della Lingua (TAL). L'applicazione di metodi e tecniche di TAL alle fonti primarie alla base degli studi in ambito umanistico rappresenta oggi una sfida aperta. Metodi e strumenti sviluppati per il trattamento delle lingue moderne non possono essere applicati come tali a varietà storiche: in questo ambito, testi riconducibili a diverse fasi evolutive di una lingua, così come testi coevi e anche lo stesso testo, presentano un'alta variabilità, che va dalla realizzazione grafica al livello morfologico, sintattico e semantico-lessicale. Il progetto TrAVaSI (*Trattamento Automatico delle Varietà Storiche di Italiano*), nato dalla collaborazione tra l'Istituto di Linguistica Computazionale del CNR e l'Accademia della Crusca e finanziato dalla Regione Toscana, si propone di costruire risorse e strumenti per il trattamento automatico di varietà storiche della lingua italiana, al momento pressoché inesistenti nel panorama nazionale e internazionale. L'obiettivo finale è la creazione dei presupposti per la messa a punto di modalità avanzate di fruizione dei contenuti culturali convogliati da varietà storiche della lingua, anche da parte del pubblico dei non addetti ai lavori. Oggetto di questo contributo è il processo di lemmatizzazione, che consiste nel ricondurre ogni parola del testo al relativo esponente lessicale o lemma. Si tratta di un livello di annotazione che presuppone la classificazione grammaticale della parola nel contesto specifico (annotazione morfo-sintattica) e che crea i presupposti per ricerche che astraggono da variazioni di natura grafica e morfologica (riguardante sia la morfologia flessionale sia quella derivazionale), e sono in grado di trattare forme sconosciute, riconducibili ad esempio voci cadute in disuso. Verranno analizzati gli effetti di due strategie di specializzazione del processo di lemmatizzazione per il trattamento di varietà storiche di italiano, operanti a livello dei dizionari utilizzati e del modello di annotazione per lemma.

2. Metodo

2.1. Il corpus

Il punto di partenza è il corpus del *Vocabolario Dinamico dell'Italiano Moderno* (VoDIM, Marazzini e Maconi, 2018), che raccoglie circa 20 milioni di occorrenze riferite a testi postunitari appartenenti a diversi domini (arte, scienza, politica, cucina) e differenziati diamesicamente. Dal VoDIM è stato selezionato un sottocorpus di circa 21.000 token, estratti da porzioni di testi in prosa di dimensioni simili (circa 3000 token ognuna), appartenenti a sette dei domini diversi (arte, cucina, diritto, giornali, letteratura, paraletteratura, scienze), diacronicamente bilanciati per avere la massima copertura possibile del periodo

[1] Per approfondimenti, cfr. anche Biffi (2016, pp. 276-77 e nota 24).

postunitario. La tabella 1 mostra la ripartizione per intervalli cronologici, che ricalcano grossomodo la suddivisione in periodi proposta per il DiaCORIS (Onelli *et al*., 2006) e successivamente per il LIS (*Lessico Italiano Scritto*)[1], a cui viene aggiunto un sesto periodo comprendente testi di italiano contemporaneo:

Tabella 1. Suddivisione del sottocorpus selezionato per intervalli cronologici

INTERVALLI TEMPORALI		N. TESTI	N. TOKEN	N. PAROLE	N. FRASI
1	1861-1900	6	3828	4111	193
2	1901-1922	5	3161	3342	173
3	1923-1945	5	2982	3174	136
4	1946-1967	5	3810	4073	141
5	1968-2001	6	3340	3546	144
6	2002-oggi	19	3503	3826	108
TOTALE		46	20624	22072	895

La tabella 2 mostra la distribuzione dei testi per domini, mantenendo il riferimento al loro periodo di pubblicazione. Il sottocorpus così definito è stato annotato morfo-sintatticamente e lemmatizzato tramite Stanza (Qi *et al*., 2020), una catena multilingue di analisi neurale basata sullo standard di rappresentazione delle *Universal Dependencies*[2] (UD, De Marneffe *et al*. 2021). È stato utilizzato il modello per l'italiano contemporaneo addestrato sulla combinazione di tutte le treebank UD disponibili per l'italiano, che coprono diversi generi testuali e varietà d'uso della lingua.

Tabella 2. Suddivisione del sottocorpus per domini

DOMINI	INTERVALLI TEMPORALI		N. TESTI	N. TOKEN	N. PAROLE	N. FRASI
arte	2-6	1902-2009	5	3225	3436	110
cucina	1-3	1871-1927	5	3071	3275	157
diritto	5-6	2002-2016	17	2552	2812	64
giornali	1-5	1867-1996	4	2572	2791	103
letteratura	1-5	1881-1982	5	3246	3391	209
paraletteratura	1-4	1892-1954	5	3085	3252	142
scienza	1-6	1864-2015	5	2873	3115	110
TOTALE			46	20624	22072	895

[2] https://universaldependencies.org/

Il testo annotato automaticamente è stato poi sottoposto a una fase di revisione manuale, che ha comportato non soltanto la correzione degli errori, ma anche l'adattamento di talune annotazioni (le categorie grammaticali) ai criteri definiti per gli UD corpora e alle specializzazioni definite per la lingua italiana. Ciò ha permesso di creare una versione gold del sottocorpus annotato, utile sia all'addestramento sia alla valutazione dei risultati conseguiti con diverse configurazioni. Il campione annotato e manualmente rivisto è stato poi suddiviso in due parti: circa l'80% è stato utilizzato per il riaddestramento dei modelli di Stanza; il 20% è stato utilizzato per la valutazione. Il corpus di riaddestramento è stato combinato con ISDT, il più esteso training corpus UD rappresentativo dell'italiano scritto contemporaneo. La valutazione ha riguardato il test corpus estratto dal VoDIM.

*Tabella 3. Suddivisione del corpus per l'addestramento e la valutazio*ne

	TRAIN	TEST
n. token ISDT	260173	
n. token VoDIM	16250	4374
totale	276423	

2.2. Lessici morfologici

I lessici morfologici possono risultare una alternativa utile per incrementare l'efficacia del processo di lemmatizzazione. Il modulo di lemmatizzazione di Stanza utilizza un dizionario costruito a partire dal corpus di addestramento (Qi *et al*., 2018).

Tabella 4. Dimensione dei lessici morfologici

LESSICO	N. FORME	N. LEMMI
VoDIM	199091	78754
SPM	44913	17404

Attingendo da due diverse fonti, sono stati creati due lessici morfologici da testi italiani del periodo '800-'900, da usare per l'integrazione del lessico base di Stanza. In entrambi i casi i lessici sono stati costruiti a partire da corpora annotati: l'uno dall'intero corpus del VoDIM, lemmatizzato automaticamente; l'altro dal corpus dei "Periodici Milanesi"[3] (De Stefanis Ciccone *et al*., 1984, in sigla SPM), riferito alla stampa periodica milanese di primo Ottocento, la cui

[3] L'archivio è consultabile online tramite gli *Scaffali digitali* dell'Accademia della Crusca (www.accademiadellacrusca.it).

annotazione è stata condotta manualmente. La tabella 4 riporta le dimensioni effettive dei due lessici.

3. Risultati della lemmatizzazione

Sono stati condotti diversi esperimenti, finalizzati alla valutazione dell'impatto delle due strategie di specializzazione definite per l'annotazione di varietà storiche di italiano, i cui risultati sono riportati in tabella 5. La colonna "Base" riporta l'accuratezza raggiunta con il modello per l'italiano moderno, le colonne "Base+Lessico" registrano i risultati conseguiti con l'estensione del lessico morfologico e la colonna "L" i risultati conseguiti con il modello riaddestrato sul corpus ISDT, esteso con il campione annotato del VODIM. Con i lessici morfologici estratti da corpora coevi si registra un guadagno rispetto al modello di base, anche se limitatamente al periodo 1_2 e ai domini arte e cucina. È interessante notare che nel caso del lessico VODIM, costruito sulla base di annotazione non rivista, si registra un decremento solo in due domini (cucina e diritto). Un più diffuso incremento di accuratezza è osservabile – anche se in media di un solo punto percentuale – nel modello derivato dal riaddestramento. Accorpando in 3 fasce i 6 periodi, tutti i modelli hanno il maggior rendimento nella fascia 3_4 rispetto alla fascia 5_6; per la fascia 1_2 si nota un discreto incremento delle prestazioni tra il modello di base e gli altri modelli. Sempre il modello di base ha incontrato più difficoltà con i testi dei domini "cucina" e "arte". Il dominio con i risultati migliori, che contiene tuttavia solo testi contemporanei, è "diritto".

Tabella 5. Accuratezza delle diverse strategie di lemmatizzazione

			BASE + LESSICO		
		BASE	SPM	VODIM	L
PERIODI	1_2	95%	96%	96%	97%
	3_4	98%	98%	98%	99%
	5_6	97%	97%	97%	99%
DOMINI	arte	95%	96%	96%	96%
	cucina	94%	97%	96%	98%
	diritto	99%	99%	98%	99%
	giornali	98%	98%	98%	99%
	letteratura	96%	96%	96%	98%
	paraletteratura	97%	97%	97%	98%
	scienza	99%	99%	99%	99%

La tabella 6 mostra l'accuratezza delle diverse strategie rispetto alle parole fuori vocabolario, cioè i vocaboli effettivamente sconosciuti (non compresi né nelle

treebank usate per l'addestramento né nei lessici morfologici sviluppati). Si nota per tutti i modelli uno scarto netto, in media di oltre il 10%, tra i risultati raggiunti nella corretta lemmatizzazione delle forme sconosciute rispetto al lessico intero, a eccezione del modello riaddestrato, il cui indice di accuratezza cala in media di un solo punto percentuale. Tali risultati vanno rapportati, considerando la piccola dimensione del corpus, a un campione di parole sconosciute molto esiguo (tra 6 e 35, se riferite ai periodi, tra 1 e 16 considerando i domini). I modelli con lessico esteso raggiungono valori inferiori al modello di base, soprattutto per quanto riguarda il VoDIM, che registra valori inferiori agli altri nella fascia 5_6, e ai testi del dominio "diritto".

Tabella 6. Accuratezza della lemmatizzazione delle parole fuori vocabolario

			BASE + LESSICO		
		BASE	SPM	VoDIM	L
PERIODI	1_2	80%	80%	80%	100%
	3_4	100%	94%	97%	100%
	5_6	84%	84%	77%	94%
DOMINI	arte	60%	60%	60%	100%
	cucina	100%	100%	100%	100%
	diritto	67%	67%	56%	89%
	giornali	100%	100%	100%	100%
	letteratura	85%	85%	77%	92%
	paraletteratura	100%	88%	94%	100%
	scienza	100%	100%	100%	100%

L'indice di accuratezza considera anche una serie di differenze che non sono errori effettivi, ma varianti dovute a diverse scelte di lemmatizzazione di talune forme: per esempio, le varianti definibili sincroniche, come i casi di regolarizzazione dei nomi alterati e derivati (*riccioletto* > *ricciolo*). Un altro tipo sono le varianti diacroniche, presenti soltanto nel lessico di SPM: si tratta di una manciata di allotropie per le quali solo uno degli allotropi viene assunto al rango di lemma, per esempio nelle alternanze protoniche del tipo *pulito/polito*, *eguale/uguale* ecc. Nelle altre configurazioni, tali varianti sono presenti ma sono state lemmatizzate separatamente. Escludendo queste da un'analisi qualitativa degli errori di lemmatizzazione, è stato possibile definire otto possibili categorie (tabella 7). La più frequente riguarda l'errata lemmatizzazione dei pronomi clitici, che incidono su oltre il 30% degli errori totali; tale valore scende a circa il 15% dei casi nel modello riaddestrato. Su una media superiore al 20% si attestano gli errori dovuti ad ambiguità, a loro volta suddivisi in ambiguità tra

parti del discorso diverse e ambiguità riferibili alla stessa classe. Il primo tipo interessa– per oltre il 70% dei casi – la sovraestensione del verbo a scapito del nome, nei contesti in cui la specifica forma genera ambiguità tra aggettivo e participio passato (*essiccato* > *essiccare*), oppure nei casi in cui lo scambio riguarda l'omonimia tra sostantivo e participio (*provvista* > *provvedere*). Sempre relativi alle ambiguità tra classi diverse, i restanti errori riguardano il fenomeno contrario, l'espansione di aggettivo e sostantivo a discapito del verbo (p.e. *distesa*, interpretato erroneamente come nome, anziché come participio di *distendere*), e in minima parte la confusione tra le classi nominali (p.e. *denigratore* in luogo di *denigratorio*). Molto meno rilevante il secondo tipo (2% dei casi), che riguarda l'ambiguità tra forme della stessa categoria; ciò avviene massimamente tra verbi (p.e. posto lemmatizzato con *postare* piuttosto che *porre*). Maggiormente incidenti sono i tre tipi che riguardano tratti grafici o soprasegmentali, le errate lemmatizzazioni per la presenza di maiuscole (7%), di sigle o acronimi (3%), di tratti paragrafematici come accenti, apostrofi o segni di punteggiatura (11%). Un errore che riguarda il modello di base è la mancata lemmatizzazione di alcune forme, errore in parte presente negli altri modelli con lessico integrato, ma assente in quello riaddestrato. Vi sono gli errori legati alla capacità di predizione del modello, che coinvolgono diverse classi grammaticali: confusioni tra le classi nominali (p.e. *carne* > *carno*), oppure, per i verbi, il mancato riconoscimento di forme irregolari o rare, che può condurre alla generazione di lemmi inesistenti (*devolvere* > *devolurare*). La distribuzione degli errori dei modelli con l'estensione del lessico sono abbastanza allineate a quelle del modello base. Ciò non vale nel caso del modello riaddestrato, che presenta una minore incidenza in numerose classi di errori (clitici, grafici e paragrafematici), ma mantiene dei livelli elevati, superiori agli altri modelli, nei casi di ambiguità e negli errori di predizione del lemma.

Tabella 7. Tipologie di errore

		BASE + LESSICO		
	BASE	SPM	VoDIM	L
ambiguità intracategoriali	1%	1%	4%	1%
ambiguità extracategoriali	13%	22%	16%	36%
maiuscole	7%	9%	8%	3%
sigle	3%	4%	2%	4%
accenti, apostrofi, segni interpuntivi	12%	13%	12%	6%
clitici	37%	37%	41%	13%
assenza del lemma	7%	2%	1%	0%
errori di predizione	20%	12%	16%	37%

4. Conclusioni

Gli esperimenti finora condotti hanno dato in sostanza esiti piuttosto positivi. Sono molte ancora le sfide aperte. Rimane insoluto il problema del trattamento delle varianti formali dello stesso lemma (cfr. § 3). Si è scelto di mantenere un basso livello di astrazione, lemmatizzando le forme separatamente da un eventuale lemma di riferimento, per evitare il rischio di normalizzazioni eccessive e di notevoli diffrazioni tra il modello riaddestrato e quelli di base. L'attuale aderenza tra le forme e i lemmi è un primo stadio: verranno pensate delle strategie per estrarre dai corpora annotati le varianti e le allotropie, raggruppandole sotto un unico lemma, considerabile come "iperlemma". Sulla base del metodo esposto, sono in corso diversi esperimenti per l'addestramento dei modelli in relazione a testi più lontani nel tempo (prosa e poesia).

Riferimenti bibliografici

Biffi M. (2016), Progettare il corpus per il vocabolario postunitario, in L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori, in Marazzini C., Maconi L., *Atti della "Piazza delle Lingue" dell'Accademia della Crusca*, edizione 2014, Firenze, Accademia della Crusca, pp. 259-80.

Bosco C., Simonetta M. e Maria S. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In 7th Linguistic Annotation Workshop and Interoperability with Discourse (pp. 61-69). The Association for Computational Linguistics.

De Marneffe M. C., Manning C. D., Nivre J. e Zeman, D. (2021). Universal dependencies. Computational linguistics, 47(2), 255-308.

De Stefanis Ciccone S., Bonomi I., e Masini A. (1984), *La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze*, Pisa, Giardini.

Marazzini C. e Maconi L. (2018). *Il Vocabolario dinamico dell'italiano moderno rispetto ai linguaggi settoriali*. Proposta di voce lessicografica per il redigendo VoDIM, Italiano digitale, 7(4): 101-120.

Onelli C., Proietti D., Seidenari C. e Tamburini F. (2006), The DiaCORIS project: a diachronic corpus of written Italian, *5th Conference on Language Resources and Evaluation* (LREC2006).

Qi P., Dozat T., Zhang Y., e Manning C.D. (2018), Universal dependency parsing from scratch, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Qi P., Zhang I., Zhang Y., Bolton J., e Manning C. D. (2020), Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Nodi discorsivi e processi di progettazione condivisa: Milano e Torino

Marco Ferracci[1], Germana Scepi[1], Maria Spano[1], Michelangelo Misuraca[2]

[1]Università di Napoli Federico II - marco.ferracci@unina.it scepi@unina.it maria.spano@unina.it
[2]Università delle Calabria - michelangelo.misuraca@unical.it

Abstract
The aim of the paper is to show the usefulness of applying the Discourse Network Analysis (DNA) in investigating policy making processes. At this aim we focus on the codesign concept in Welfare and choose Turin and Milan as case studies. The database is composed of municipal or territorial acts observed form 2016 to 2021.They are extracted by municipal websites following a double strategy of research, qualitative and key oriented. The analysis is conducted by means of two software: "DNA analyzer" and Ucinet for graphical representations. The results show two similar definitions of codesign in the two cities.

Keywords: codesign, governance, discourse network analysis, Italy

Riassunto
L'obiettivo di questo lavoro è di mostrare le potenzialità della Discourse Network Analysis (DNA) nell'ambito dei processi di policy making. A tal fine è stato scelto di analizzare le "pratiche discorsive" sul concetto di coprogettazione nel campo del Welfare a Milano e Torino, Il database è composto di una serie di atti comunali o territoriali, estratti da database municipali, nell'intervallo temporale che va dal 2016 al 2021. La scelta ha seguito un duplice criterio di ricerca qualitativo e key oriented. Per l'analisi è stato utilizzato il software "DNA analyzer" mentre per la rappresentazione grafica, Ucinet. I risultati mostrano due approcci simili nella definizione della coprogettazione nelle due città.

Keywords: coprogrettazione, governance, discourse network analysis, Italia

1. Introduzione

Il termine coprogettazione è caratterizzato da forti ambivalenze concettuali e implementative nel campo del Welfare. Infatti, oltre agli aspetti innovativi e positivi riguardanti l'azione pubblica, la rivitalizzazione della società civile e più in generale di spazi di partecipazione, vi sono molti aspetti critici. In Italia in

particolare alcuni cambiamenti normativi, sia relativi all'autonomia degli enti locali, sia alla possibilità di esternalizzare specifici compiti da parte degli organi locali dello stato (per citarne alcune la legge 328/2000, la riforma costituzionale del 2001, e il "codice del terzo settore" Dlgs. N.117 del 2017), hanno messo in luce come "nuovi attori" debbano essere presi in considerazione per una definizione efficace del Welfare (Ranci 1999). Diventa quindi fondamentale comprendere i sistemi di governance e i meccanismi relazionali che governano la città. Per studiare questi fenomeni è stata scelta un tipo di analisi qualitativa del discorso. Si farà riferimento alla Discourse Network Analysis (DNA) come proposto da Philip Leifeld (2017). Tradizionalmente questo strumento viene utilizzato per analizzare le pratiche discorsive legate al policy making. In questo lavoro si farà riferimento alle differenze discorsive che accompagnano i processi di coprogettazione dei servizi Welfare nelle città di Milano e Torino dal 2016 al 2021. L'obiettivo del lavoro è di mostrare le potenzialità della DNA per l'analisi dei processi discorsivi sul tema della coprogettazione.

2. Framework Teorico

La ricerca fa riferimento alla teoria dei policy network di matrice costruttivista (Bevir e Rhodes 2007) e della Discourse Network Analysis (Leifeld 2017). L' analisi DNA parte dall'assunto teorico che il discorso politico sia un fenomeno reticolare. Le affermazioni degli attori sono interdipendenti, in quanto in grado di influenzare o apprendere dagli altri attori nel dibattito. Come affermato da Leifeld l'obiettivo principale di questo tipo di analisi è descrivere e trovare le ricorsività nella struttura del discorso, in molti casi politico, attraverso gli strumenti della social network analisys. L'analisi si basa su quattro tipi di variabili: attori, ovvero gli individui singoli o organizzativi che producono le affermazioni; i concetti, ovvero la rappresentazione astratta dei contenuti discussi; il tempo, come variabile discreta o continua nel quale si svolgono i dibattiti; in ultimo, l' accordo o disaccordo degli attori rispetto a una determinata affermazione. Alcuni esempi applicativi riguardando il dibattito intorno alle politiche pensionistiche in Germania (Leifeld 2013), sul dibattito europeo riguardo i brevetti di alcuni software (Leifeld e Haunnass 2010), sul prezzo minimo degli alcolici (Fergie *et al*. 2018) o le food policies in alcune città europee(Filippini *et al*. 2019). La DNA è una alternativa analitica valida alle classiche prospettiva di analisi del discorso. Come affermato da Leifeld e Haunns (2010) questa tecnica ha il vantaggio di includere variabili centrate sia sugli attori del discorso (come nella tradizione dell'advocacy coalition framework, della discourse equilibrium theory, policy paradigm ecc…);sia sul contenuto del discorso (come proposto dalla critical discourse analysis, semantic network analysis). Inoltre, la DNA ha il vantaggio di rendere osservabili le strutture discorsive in termini di formalizzazione attraverso gli strumenti della social

network analysis andando a incrementare le capacità analitiche per lo studio delle strutture di potere. In entrambi gli approcci sopra menzionati l'attenzione è posta sui processi di costruzione discorsiva/interazionale del policy making. La costruzione degli spazi discorsivi tra pubblico e privato è un elemento centrale per comprendere la strutturazione e l'implementazione dei servizi di Welfare (Gaynor 2011).

3. Dati e Metodi

Il lavoro si basa sull'analisi qualitativa del discorso, in particolare sulla Discourse Network Analysis (Leifeld 2017). La collezione di documenti è composta di una serie di atti comunali o degli ambiti territoriali (delibere comunali, Piani di zona, atti di indirizzo programmatico) scelti seguendo un duplice criterio: il primo qualitativo volto ad identificare i documenti maggiormente significativi nella definizione del concetto di coprogettazione; il secondo maggiormente quantitativo condotto attraverso una ricerca per keywords sui siti dei due comuni. Le chiavi di ricerca utilizzate sono: "coprogettazione", "co-programmazione", "coprogettante", "e "progettazione condivisa". All'interno dei rispettivi siti web comunali sono stati applicati dei filtri in modo da selezionare solo documenti che riguardassero i temi del Welfare dal 2016 al 2021.[1] Il database si compone di 24 documenti dei 44 inziali per Milano e 27 documenti dei 74 iniziali per Torino. Ogni documento è stato letto cercando le chiavi "coprogettazione" , "co-programmazione", "coprogettante" e "progettazione condivisa" all'interno del testo. Durante la prima fase di analisi sono stati eliminati i documenti che contenevano informazioni ridondanti. Dopo di che gli estratti ritenuti interessanti sono stati riportati ne lsoftware Java open access Discourse Network Analyzer creato dal Paul Leifeld. L'analisi è stata condotta annotando manualmente le singole affermazioni, per un totale di 80 annotazioni, sulla base delle variabili: attore e concetto nel range temporale scelto 2016 al 2021. I risultati dell'analisi sono stati poi esportati analizzati e visualizzati graficamente su UCINET utilizzando NETdraw. Sono state inoltre estrapolate le parole più utilizzate per caratterizzare i diversi livelli analitici attraverso delle nuvole di parole create grazie a word cloud generator di cui verranno presentate solo le parole principali per questioni di spazio.[2]

[1] https://www.comune.milano.it/; http://www.comune.torino.it/
[2] https://www.visual-thesaurus.com/wordcloud_generator_italiano.php

4. Risultati

Al fine di analizzare le suddette dinamiche verranno utilizzate le categorie di coprogettazione proposte da Rossi e Colombo (2019). Gli autori identificano due tipi di domini analitici rispetto al tema, che tuttavia possono ibridarsi creandone un terzo: coprogettazione come *logica d'azione organizzativa* e che riguarda la volontà di rendere maggiormente democratico il sistema di gestione dei servizi, distribuendo responsabilità e ruoli. Concetto legato quindi più alla partecipazione degli attori che alla definizione normativa apriori delle modalità di azione. Coprogettazione come *dispositivo amministrativo* atto a regolare i rapporti contrattuali e quindi a definire il frame normativo per l'implementazione di un servizio. Infine viene riportato un *modello ibrido* che riguarda il passaggio da un livello analitico all'altro. Questi tre livelli analitici sono stati chiariti nelle prime due colonne della tabella 1.

Tabella 1. Coprogettazione livelli analitici, dimensioni, concetti e analisi, nostra elaborazione Rossi e Colombo (2019)

Coprogettazione			**Analisi**
Dimensioni	**Livelli analitici**	**Concetti**	**Parole caratterizzanti i livelli analitici**
1) promuovere il coinvolgimento attivo della pluralità di attori in modo non formalizzato 2) possibilità di accordi informali 3) può nascere anche senza precedenti relazioni formalizzate	**Logica d'azione organizzativa**: quell'insieme di attività focalizzate sia sulla costruzione di un intervento, sia sulla definizione dei meccanismi di governance tra gli attori coinvolti	1) promuovere il coinvolgimento degli attori 2) esistenza di un accordo informale	**Milano**: progettuale, servizio, intervento, comunale, garantire, risorse, avviso, progetti, soggetti, attività, partecipazione, territorio, amministrazione, pubblico **Torino**: partner, città, autonomia, pubblica, accompagnamento, funzione, soggetti, persone, progettuale, partecipazione, proposta, collaborazione, percorsi
1) formalizzazione di un documento normativo che definisca chiaramente vincoli e possibilità	**Dispositivo amministrativo**: regolare i rapporti contrattuali e quindi a definire il frame normativo	1) chiara definizione dei ruoli/risorse e responsabilità 2) creazione framework	**Milano**: proposta, partner, comunale, selezione, enti, avviso, attività, fase, progetti, intervento, soggetti, amministrazione, valore, realizzazione,

2) definizione temporalmente antecedente del sistema di governance 3) chiara definizione dei ruoli/risorse e responsabilità	per l'implementazione di un servizio	normativo	valutazione, progettazione, procedura, interventi **Torino**: piano, progettuali, progetto, sistema, inclusione, accompagnamento, area, azioni, attività, fase, sviluppo, accordo, collaborazione, modalità, cittadino, realizzazione, partner
1) passaggio da una condizione informale a una formalizzata 2) fasi temporalmente successive 3) passaggio da una definizione non formalizzata a una formalizzata dei ruoli	**Misto:** transizione da un livello analitico all'altro. Il raccordo tra queste due interpretazioni si trova, evidentemente, nella possibilità che tale dispositivo amministrativo si «appoggi» su una rete preesistente.	1) passaggio da una rete informale a una formalizzata 2) rafforzamento di una rete esistente verso un framework maggiormente regolato	**Milano**: consolidata, marginalità, enti, sociali, responsabilità, città, filantropiche, pubblico, azione, collegamento, rete, volontariato, esperienze, terzo, settore **Torino**: gestione, amministrazione, risorse, interventi, campo, politiche, definizione, istituzionale, cofinanziamento, potenziare, servizi, innovazione, necessità, sperimentazione, disposizione, soggetti, regia, integrazione, privato, implementare

Di seguito sono riportati i concetti operativizzati utilizzati per l'analisi: Con i concetti di *creazione di un framework normativo e chiara definizione dei ruoli/risorse e responsabilità* si fa riferimento a quello che Rossi e Colombo (2019) chiamano dispositivo amministrativo. Il primo riguarda la definizione generale delle condizioni contrattuali mentre il secondo la chiara e formalizzata divisione dei ruoli, risorse e responsabilità tra gli attori. Con *passaggio da una rete informale a una formalizzata e rafforzamento di una rete esistente verso un framework maggiormente regolato* si fa rifermento alla dimensione mista. Il primo fa riferimento alla trasformazione da rete informale a formale, mentre il secondo può riguardare anche reti con un basso grado di formalizzazione. *Promuovere il coinvolgimento degli attori* richiama la dimensione dell'azione organizzativa come creazione di spazi o opportunità per includere nuovi attori nei processi di progettazione; in ultimo *esistenza di un accordo informale* fa riferimento a una logica di azione creata attraverso accordi informali, come per esempio il prestito di spazi e strumenti, per promuovere l'azione degli attori. Per questioni di spazio verranno riportati solo i principali risultati dell'analisi. La

tabella 2 mostra le differenze nella definizione di coprogettazione nel campo del welfare nelle due città nel periodo 2016-2021. Entrando nel merito dei risultati si può notare come la categoria prevalente per entrambe le città sia legata alle strategia per allargare la partecipazione. Infatti in entrambi i casi questa è la categoria che registra il maggior numero di annotazioni. Anche la creazione delle condizioni normative per la coprogettazione mostra valori elevati, tuttavia se si guarda alla colonna uno, chiara definizione dei ruoli, risorse e responsabilità degli attori, Milano risulta porre meno attenzione su questa dimensione rispetto a Torino. Riguardo il livello analitico misto, che fa riferimento ai concetti nella quarta e quinta colonna, questi nodi discorsivi registrano minore attenzione da parte di entrambe le città. In modo particolare Torino presta minore attenzione al dibattito sulla formalizzazione.

Tabella 2. Risultati dell'analisi DNA

	Chiara definizione dei ruoli/risorse e responsabilità	Creazione di un framework normativo	Promuovere il coinvolgimento degli attori	Esistenza di un accordo informale	Rafforzamento di una rete esistente verso un framework maggiormente regolato	Passaggio da una rete informale a una formalizzata
Milano	3	10	14	1	5	4
Torino	9	14	13	1	5	1

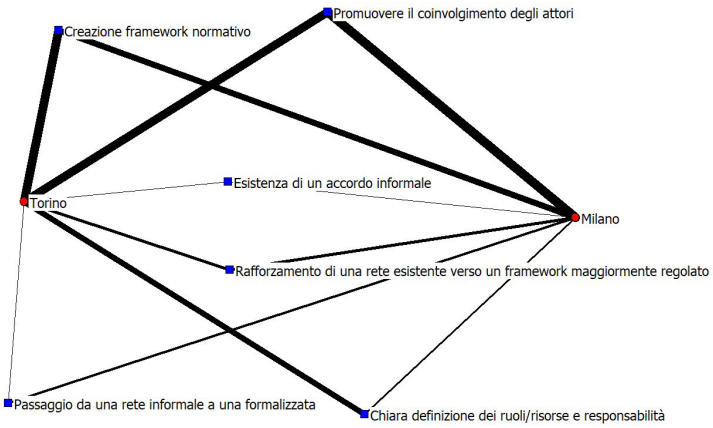


*Figura 1. network 2-mode attori*concetti. Ai legami è stata applicato il peso del legame riportato nella tabella 1.*

La figura 1 riporta il grafo a 2-modalità (attore*concetto) come trasposizione grafica della tabella 2. Lo spessore dei legami fra le città e i concetti esprime il peso del legame, ovvero il numero di annotazioni riscontrate per quel legame. La tabella 1 nella colonna analisi mostra le parole caratterizzanti dei differenti livelli analitici. Per entrambe le città la logica di azione è legata a parole riguardanti la partecipazione e la collaborazione degli enti nella progettazione dei servizi. Nel secondo livello analitico invece le parole che ricorrono sono legate alla definizione di un percorso condiviso scandito da fasi. Nel livello misto invece le parole sono maggiormente legate a strategie di consolidamento e o definizione di reti di intervento.

5. Conclusioni

L'obiettivo del lavoro è di mostrare le potenzialità della DNA per l'analisi dei processi discorsivi sul tema della coprogettazione nelle città di Milano e Torino riguardante il Welfare tar 2016 e il 2021. Per la definizione delle dimensioni è stato utilizzato lo schema analitico proposto da Rossi e Colombo (2019). Il lavoro mostra due pratiche discorsive simili per le due città. Infatti, la categoria concettuale che registra più annotazioni è *Promuovere il coinvolgimento degli attori.* Le pratiche discorsive quindi si concentrano nel delineare spazi di partecipazione e coinvolgimento multi attore. I principali vantaggi dell'utilizzo della DNA è che le pratiche discorsive vengono concepite come una rete composta di attori e concetti, graficamente rappresentabili e analizzabili attraverso gli strumenti della Social Network Analysis. L'analisi quindi restituisce diversi livelli analitici che possono essere studiati insieme, come nel caso dell'analisi condotta, o separatamente, ad esempio attraverso network one-mode per riga o per colonna.

Bibliografia

Bevir M. e Rhodes R.A.W. (2007). Decentred Theory, Change and Network Governance. In Sørensen, E. e Torfing, J. a cura di, *Theories of democratic network governance.* Palgrame Macmillan.

Fergie G., Leifeld P., Hawkins B. e Hilton S. (2018). Mapping discourse coalitions in the minimum unit pricing for alcohol debate: a discourse network analysis of UK newspaper coverage. *Policy Case Study*, 114(4): 741-753.

Filippini R., Mazzocchi C. e Corsi S. (2019). The contribution of Urban Food Policies toward food security in developing and developed countries: A network analysis approach. *Sustainable Cities and Society*, 47: 1-13.

Gaynor N. (2011). Associations, Deliberation, and Democracy: The Case of Ireland's Social Partnership. *Politics and Society*, 39(4): 497-519.

Leifeld P. (2013). Reconceptualizing Major Policy Change in the Advocacy Coalition Framework: A Discourse Network Analysis of German Pension. *The Policy Studies Journal*, 41(1): 169-198.

Leifeld P. (2017). *Discourse network analysis: policy debates as dynamic networks.* In

Victor, J. N. e Montgomery M. N. a cura di, *The oxford Handbook of political networks*. Oxford University press.
Leifeld P. e Haunnss S. (2010). A comparison between political claims analysis and discourse network analysis: the case of software patents in the European Union, Preprints of the Max Planck Institute for Research on Collective Goods.
Rossi P. e Colombo M. (2019). Non sarà un'avventura? L'innovazione delle partnership pubblico- privato e la co-progettazione dei servizi di welfare sociale. *Stato e mercato*, 3/2019: 411-447.
Ranci C. (1999). *Oltre il welfare state. Terzo settore, nuove solidarietà e trasformazioni del welfare*. Bologna, Il Mulino.

Text Mining e Salute Organizzativa: il caso di Roma Capitale

Rosa Ferri[1], Valentina Bua[1], Marzia Curia[1], Alessia D'Avack[2], Antonino Firetto[2], Federica Mauro[1], Francesca Greco[1]

[1]Sapienza Università di Roma – rosa.ferri@uniroma1.it, valentina.bua@uniroma1.it , marzia.curia@hotmail.it, federicamauro4683@gmail.com, francesca.greco@uniroma1.it
[2]Roma Capitale – alessia.davack@comune.roma.it, antonino.firetto@comune.roma.it

Abstract
Our paper's aim is to describe how the Emotional Text Mining (ETM) has been used to evaluate the impact of an Organisational Health promotion intervention on City of Rome's employees. Together with Università Sapienza, we followed 140 city employees - serving as Organizational Health representatives- who took part in a training program, through an established time frame (including a follow-up evaluation 12 months after the end of the program). Through the analysis of the participants' project work and focus groups we were able to demonstrate that the program was successful in promoting cultural change, enabling organizational and personal empowerment.

Keywords: Emotional Text Mining; Organizational Health; assessment; training program

Riassunto
Questo contributo illustra l'utilizzo dell'Emotional Text Mining (ETM) per la valutazione di un intervento di promozione della salute organizzativa realizzato a Roma Capitale, in collaborazione con l'Università Sapienza di Roma. Nello specifico l'ETM è stata utilizzata per la valutazione di processo, di esito e per un follow up ad 1 anno di un percorso formativo destinato a 140 dipendenti deputati alla funzione di Referenti della Salute Organizzativa (RSO). L'analisi ha coinvolto i testi dei project work prodotti dai formandi durante il corso e i trascritti dei focus group realizzati con i partecipanti prima dell'inizio del corso e ad un anno dal termine. La valutazione ha evidenziato come il corso abbia favorito un cambiamento culturale, attivando e mantenendo nel tempo processi di empowerment personale e organizzativo.

Parole chiave: Emotional Text Mining; Salute Organizzativa; valutazione; formazione

1. Introduzione

Questo lavoro presenta l'utilizzo delle tecniche di text mining per il monitoraggio e la valutazione del processo formativo volto a promuovere la salute organizzativa realizzato da Roma Capitale. Un processo che ha inteso attivare e supportare il cambiamento della *cultura locale* (Carli e Pagano, 2008; Cordella *et al*., 2014), capace di assicurare un adeguato grado di benessere fisico e psicologico dei lavoratori, attraverso l'acquisizione di nuove conoscenze e il cambiamento del modo di rappresentare il processo di inclusione delle persone con diversa abilità. In particolare, Roma Capitale, in collaborazione con l'Università di Roma "Sapienza", ha realizzato un percorso formativo destinato a 140 dipendenti deputati alla funzione di Referenti della Salute Organizzativa (RSO), con l'obiettivo di promuovere la salute nell'organizzazione favorendo l'attivazione di processi d'inclusione del personale dipendente con disabilità.

Il corso, realizzato tra settembre 2019 e luglio 2020, ha avuto l'obiettivo di trasmettere agli RSO competenze che consentano loro di svolgere attività di promozione dei processi di transizione lavorativa, accoglienza, collocazione dei nuovi assunti e dipendenti in mobilità interna, con particolare attenzione all'inserimento dei lavoratori con disabilità. Il processo formativo è stato pensato alternando l'apprendimento teorico all'apprendimento in situazione. Attraverso una metodologia "learning by doing" i formandi hanno potuto mettere in pratica ed implementare le conoscenze acquisite durante il corso per fare ricerca o progettare interventi nei contesti reali, costruendo progetti di promozione della salute organizzativa.

Al fine di effettuare la valutazione di processo e di esito della formazione, alla fine del corso e dopo un anno, si è scelto di tenere in considerazione: gli aspetti qualitativi e quantitativi, quelli oggettivi e soggettivi e la dimensione temporale dell'acquisizione delle competenze. Di conseguenza, gli strumenti di raccolta dei dati selezionati sono stati sia scritti che orali, determinando la necessità di utilizzare sia tecniche tradizionali di analisi dei dati che procedure di Natural Language Processing (NPL) capaci di rilevare il cambiamento culturale. In particolare sono stati utilizzati tre strumenti di rilevazione: il questionario a risposta multipla, per la valutazione delle conoscenze apprese, la formulazione di un progetto e i focus group.

Il cambiamento della cultura locale è stato valutato attraverso l'analisi dei trascritti dei focus group e dei project work realizzati dagli RSO. I testi sono stati analizzati con il metodo dell'Emotional Text Mining (ETM), (Greco, 2016; Greco e Polli, 2020) una metodologia di Natural Language Processing (NPL) che si colloca in una prospettiva di matrice socio- costruttivista a orientamento psicodinamico. L'ETM, attraverso un approccio sia semiotico che semantico al testo, consente di identificare tre livelli della comunicazione: il contenuto, le rappresentazioni sociali e la matrice simbolico-culturale generatrice delle

dinamiche relazionali e produttive entro i contesti organizzativi. L'ETM si basa sull'utilizzo di tecniche multivariate (analisi dei cluster e analisi delle corrispondenze) e non necessita di dizionari di riferimento o di un learning set per la classificazione dei testi. In particolare, l'ETM è stato utilizzato per analizzare i project work redatti dagli RSO ad 1 mese, 4 mesi e al termine della formazione (8 mesi), consentendo il monitoraggio di processo e la valutazione dell'esito del corso, e per l'analisi dei trascritti dei focus group realizzati prima dell'inizio del corso e a due anni, consentendo di valutare le dimensioni del cambiamento culturale che permangono nel tempo.

2. Metodologia

I project work e le trascrizioni dei focus group sono stati raccolti in due corpora distinti, che sono stati analizzati con il metodo ETM procedendo come segue per ciascun corpus.

Inizialmente i testi sono stati pretrattati eliminando le parti grafiche, i link, le e-mail e i numeri mantenendo solo il testo in minuscolo. Successivamente, sono stati calcolati tre indicatori lessicali per valutare la possibilità di processare statisticamente i dati: i token, il type-token ratio e la percentuale di hapax (Giuliano e La Rocca, 2010) e ogni corpus è stato suddiviso in frammenti di testo (unità di contesto), lemmatizzato e sono state escluse le stop word e i termini di rango alto e basso di frequenza (Bolasco, 1999) al fine di ridurre la dimensione della matrice (Greco, 2016). Sulla matrice termini-unità di contesto è stata effettuata un'analisi dei cluster con un algoritmo bisecting k-means, limitato a 20 partizioni, con l'esclusione dei testi che non presentavano almeno una co-occorrenza, ed è stato utilizzato il coseno come misura di similarità (Steinbach *et al*., 2000). Il Calinski-Harabasz, il Davies-Bouldin e il coefficiente di correlazione intraclasse sono stati calcolati per individuare la partizione ottimale e sulla matrice termini- cluster è stata effettuata un'analisi delle corrispondenze (Lebart e Salem, 1994) per identificare gli assetti simbolici caratterizzanti la cultura locale (Cordella *et al*., 2014; Greco, 2016). Infine, sulla tabella di contingenza cluster-tempo sono stati calcolati il χ^2 e i residui standardizzati per valutare il cambiamento delle rappresentazioni (cluster) nel tempo: 1 mese, 4 mesi, 8 mesi per i project work e prima del corso e dopo due anni per i focus group.

3. Risultati

Le analisi effettuate sui testi dei project work e dei focus group hanno generato 2 diversi spazi fattoriali, all'interno dei quali è stato possibile individuare le dimensioni caratterizzanti la cultura locale degli RSO durante, al termine e a due anni dall'inizio dalla formazione. Nello specifico, i risultati dell'analisi dei project work sono riassunti in tabella 1 e quelli dei focus group in tabella 2.

Tabella 1. Quadro sintetico dell'interpretazione dell'analisi multivariata dei project work (tra parentesi accanto al fattore è riportata la percentuale di inerzia spiegata e sotto l'etichetta dell'asse fattoriale il valore della coordinata del centroide del cluster)

Cluster	*Rappresentazioni*	*UC*	*UC%*	*Fattore 1 (40,5%)*	*Cultura Locale* *Fattore 2 (32,0%)*	*Fattore 3 (27,5%)*
				Premesse	*Processo*	*Contesto*
1	Progettare l'inclusione	1182	21 %	Personali -0,59	Inclusione 0,51	Relazionale -0,49
2	Inserimento organizzativo	1554	27 %	Sistemiche 0,39	Accoglienza -0,50	Relazionale -0,41
3	Funzioni del RSO	1634	29 %	Sistemiche 0,49	Inclusione 0,41	Strutturale 0,33
4	Valutazione di vincoli e risorse	1305	23 %	Personali -0,58	Accoglienza -0,40	Strutturale 0,48

UC: unità di contesto

I project work si sviluppano all'interno di uno spazio di senso definito da tre assi principali: le premesse, il processo e il contesto entro i quali si collocano quattro rappresentazioni del lavoro degli RSO: progettare l'inclusione, favorire l'inserimento organizzativo, svolgere la funzione del RSO e valutare vincoli e risorse.

Tabella 2. Quadro sintetico dell'interpretazione dell'analisi multivariata dei focus group (tra parentesi accanto al fattore è riportata la percentuale di inerzia spiegata e sotto l'etichetta dell'asse fattoriale il valore della coordinata del centroide del cluster)

Cluster	*Rappresentazioni*	*UC*	*UC%*	*Fattore 1* *27,7%* *Intervento*	*Cultura Locale* *Fattore 2* *22,4%* *Potere*	*Fattore 3* *20,5%* *Contesto*
1	Mandato	428	14%	Acontestuale 0,43	Ruolo -0,51	
4	Difficoltà	601	20%	Acontestuale 0,47	Competenza 0,62	
6	Appartenenza	343	11%	Acontestuale 0,46	Ruolo -0,44	
2	Autogestione proattiva	678	22%	Contestuale -0,40		Relazione -0,57
3	Analisi di contesto	591	19%	Contestuale -0,36		Struttura 0,38
5	Confronto con l'estraneo	415	14%	Contestuale -0,33		Struttura 0,41

UC : unità di contesto

Parallelamente, i focus group si sviluppano all'interno di uno spazio di senso definito da cinque assi principali di cui i primi tre spiegano il 70,6% dell'inerzia: l'intervento, il potere e il contesto. È interessante notare come in questo caso il primo fattore divide e distingue le rappresentazioni in due gruppi, interventi contestuali e interventi acontestuali, e che la capacità di pensare l'intervento in

termini di contesto sia uno dei prodotti del corso.
Infatti, le rappresentazioni variano significativamente nel tempo sia per quanto riguarda i project work ($\chi 2$, df= 6, p>0,01) che i focus group ($\chi 2$, df= 5, p>0,01). Come si evidenzia in tabella 3, la capacità degli RSO di pensare ad un progetto capace di promuovere l'inclusione è significativamente assente ad un mese dall'inizio del corso (Progettare l'inclusione, SE= -1,97). Inoltre, se prima dell'inizio della formazione il tema delle difficoltà è significativamente presente (Difficoltà, SE= 3,55) e quello dell'autogestione proattiva è assente (Autogestione proattiva, SE= -3,47), ad un anno dalla fine del corso la capacità di autogestione proattiva è presente (Autogestione proattiva, SE= 3,21), mentre il tema delle difficoltà assente (Difficoltà, SE= -3,28).

Tabella 3. Residui standardizzati dell'associazione dei cluster per tempo dei project work e dei focus group

Rappresentazioni	*Project work*			*Rappresentazioni*	*Focus Group*
	1 mese	*4 mesi*	*fine corso*		*Prima2 anni*
Progettare l'inclusione	-1,97	0,98	0,02	Mandato	0,91-0,84
Inserimento organizzativo	1,56	-0,95	0,15	Difficoltà	3,55-3,28
Funzioni del RSO	-0,85	0,58	-0,14	Apartenenza	-1,511,40
Valutazione di vincoli e risorse	1,12	-0,54	-0,02	Autogestione proattiva	-3,473,21
				Analisi di contesto	-0,930,86
				Confronto con l'estraneo	1,72-1,59

4. Discussione e conclusioni

L'utilizzo delle tecniche di text mining ha consentito di evidenziare come i risultati ottenuti con il corso permangano a due anni di distanza e quali assetti simbolico-culturali siano stati capaci di promuovere la salute organizzativa. Possiamo quindi ipotizzare che il valore aggiunto apportato dalla formazione è l'incremento dell'empowerment organizzativo grazie ad uno spostamento da una posizione di delega, dovuta alla difficoltà di individuare il possesso di specifiche competenze e ad intervenire, ad una posizione proattiva, in cui si percepisce la propria capacità di leggere e trasformare il contesto non connessa al ruolo, ma alla propria competenza.
La valutazione ha tracciato una chiara traiettoria di sviluppo della cultura degli RSO, evidenziando come l'attivazione dei processi di empowerment personale e organizzativo promossi dal corso sia pienamente apprezzabile ad un anno dalla conclusione. In particolare, sono emerse: la competenza a pensare secondo una dimensione progettuale, secondo cui l'inserimento lavorativo è un processo inclusivo che si dipana nel tempo e che va costantemente monitorato. In questo processo l'RSO si pone come connettore e facilitatore dell'incontro tra individuo

e contesto; la dimensione dell'appartenenza ad una rete professionale che sostiene e valorizza le competenze acquisite; la rappresentazione della disabilità non come elemento precostituito, ma da conoscere, che favorisce l'adozione di una posizione esplorativa e avalutativa volta a cogliere le risorse dell'altro piuttosto che evidenziarne le mancanze; il passaggio da una posizione passiva, caratterizzata dall'espressione di difficoltà e vissuti di impotenza, ad una rappresentazione di sé come agenti proattivi, in grado di riconoscersi competenze, assumersi funzioni attive e generare cambiamento.
Concludendo, il processo di valutazione ha permesso di cogliere i cambiamenti generati dalla formazione in una prospettiva dinamica, evidenziandone l'evoluzione nel tempo. In questo senso è possibile affermare che il percorso di formazione, nello sviluppare competenze, abbia attivato un processo di crescita dell'empowerment a livello personale e organizzativo che non si è arrestato dopo la fine del corso, ma permane ad un anno dalla conclusione.

Bibliografia

Bolasco S. (1999). *Analisi multidimensionale dei dati: Metodi, strategie e criteri d'interpretazione*. Carocci.
Carli R. e Pagano P. (a cura di) (2008). *San Lorenzo. La cultura del quartiere e i rapporti con psicologia*. Kappa.
Cordella B., Greco F. e 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.
Giuliano L. e 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. e Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management*, 51:101934.
Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod.
Steinbach M., Karypis G. e Kumar V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 400, pp. 525–526. Boston.

Lasso-based variable selection methods in text regression: the case of short texts

Marzia Freo[1], Alessandra Luati[2]

[1]European Commission, Joint Research Centre (JRC) – marzia.freo@ec.europa.eu
[2]University of Bologna – alessandra.luati@unibo.it

Abstract
Communication through websites is often characterised by short texts, made of few words, such as image captions or tweets. This paper explores the class of supervised learning methods for the analysis of short texts, as an alternative to unsupervised methods, widely employed to infer topics from structured texts. The aim is to assess the effectiveness of text data in social sciences, when they are used as explanatory variables in regression models. We compare the results obtained by several variants of lasso, screening-based methods and randomization-based models, such as sure independent screening and stability selection. Latent dirichlet allocation results are also considered as a term of comparison. Our perspective is primarily empirical and our starting point is the analysis of two real datasets, though bootstrap replications of each dataset are considered. The first case study aims at explaining price variations based on the information contained in the description of items on sale on e-commerce platforms. The second regards open questions in surveys on satisfaction ratings.

Keywords: Text Mining, Lasso, Variable Screening, Stability Selection, Latent Dirichlet Analysis.

1. Introduction

Over the past decades, new technologies and the development of online social platforms have made available to researchers a large amount of text data. Consequently, many studies have been conducted with the aim of exploiting the informative content of digital text. Currently, texts are used as data in a variety of applications in favour of social and economic insights: authorship, sentiment, nowcasting, policy uncertainty, media slant, market definition, and other topics, as it is witnessed by the review by Gentzkow et al. (2019).

This paper explores the class of supervised learning methods for the analysis of short texts, as an alternative to unsupervised methods, widely employed to infer topics from structured texts. With the focus on interpretation we investigate the

potential of text regression when it is performed by recurring to variable selection models. We compare the results obtained by several variants of lasso, screening-based methods and randomization-based models. Latent Dirichlet Allocation (LDA) results are also considered as a term of comparison.
We consider two applications. In the first case study, we afford the issue of explaining variation of prices of goods within the same category, based on the descriptive text or label provided by producers on e-commerce platforms. Solving this task may provide new insights on hedonic evaluation and on price index measurement. Moreover, whether estimated prices were made available, they would decrease informative asymmetries in the markets, by opening information to consumers. Concerning the second case study, we consider how open questions inserted in a questionnaire may be informative on the overall satisfaction ratings. Here, the main rationale is that open questions may raise opinions and point of views that could have not been planned by traditionally collected questionnaires with predefined attributes.

2. Data description and preprocessing

2.1. Case study 1

The first case study focuses on the task of pricing items available on producers' e-commerce plat-forms. Our specific aim is to model regular fashion prices within a cross-sectional comparison, as we are not primarily interested in price dynamics. We use data scraped from UK e-commerce websites of four fashion producer brands.
Before affording the issue of sparse modelling, we process some preliminary steps to reduce the dimensionality and to map raw text into a numerical matrix, the Document Term Matrix (DTM), whose ij-th element indicates the count of the j-th word or token in the i-th document. Some text preparation operations are required in order to process the data and reduce meaningless dimensionality. First, we cancel out non-words elements (like numbers, punctuations, and proper names); then very common words are collected in a stop-words list and deleted; finally, words are replaced with their roots through stemming and very uncommon stems are removed. After the pre-processing step, the DTM matrix for the knitwear dataset N = 382 and P = 229 words, the dresses dataset N = 1110 and P = 402 columns. In both datasets each document is composed, in median, by only three words, each word appears in less than 1% of documents. As a whole DTMs are highly sparse, with about 99% of empty cells.

2.2. Case study 2

The second case study is related to a section of the Tech Company Employee reviews dataset, downloaded from www.kaggle.com. Data were scraped from www.glassdoor.com. Glassdoor is a website which allows current and former

employees to anonymously review companies and also to anonymously submit and view salaries, as well as to search and apply for jobs on the same platform. The analyses are carried out for reviews about a worldwide Tech anonymous company. Each record collects information on company, location, date of the review, job, position. Moreover, it collects evaluations on a 1-to-5 scale on overall-ratings and other aspects concerning the job, along with two further open questions on pros and cons. Words deriving by the positive field, pros, are presented preceded by the prefix "p_" while words deriving by the negative significant, cons, are preceded by the prefix "c_". Our aim is to explain the overall-rating using only the text contained in pros and cons open questions fields as data.

We shall treat ratings as a response variable in text regression. Though more specific methods can be applied to interval scales and ordinal variables (see for instance Hastie et al. 2009, ch. 14), for sake of interpretation and comparison, and in order to effectively perform variable selection, we shall investigate the marginal contribution of attributes by estimating linear regression models. The motivation for a regression analysis on rating is twofold. First, variable selection methods for categorical data are not as developed as methods for continuous variables and thus often do not allow for homogeneous comparisons. Second, ratings are most analysed as quantitative variables, which makes the results of the analysis in the paper to be useful in several related applications.

After a similar text pre-processing to the one described in the previous section, the DTM for the Employees dataset is a N = 808 by P = 1135 matrix (with bigrams). Each document is composed in median by 15 words. Each word, in mean, is present in the 1.7% of documents, and the sparsity of the DTM reaches the 98.3%.

3. Design of the study

To evaluate to which extent relevant variables have been detected is a challenging task, as the true model is unknown but in simulations. In practical applications, only one dataset at a time is given and one does not know which variables are truly influential; we only may express judgments about their plausibility. To mimic the availability of several datasets, we recur to bootstrap replications and resample (five hundred times) by each same unique dataset. Indeed, the bootstrap may yield desirable perturbations similar to those of multiple data sets (Efron and Tibshirani, 1998).

Each replicated dataset is divided into a training set and a hold-out sample. For each bootstrap dataset, we first select variables over training data, using a pool of models, then we estimate a linear regression model by ordinary least squares using previously selected variables as predictors, over the hold-out sample and finally we measure the different variants of R^2. To account for possible overfitting, we present both the adjusted and predictive R^2. We remark here that

we compute the predictive R^2, not because we aim to evaluate the usefulness of selected variables for out-of-sample forecasting, which is out of our scope, but rather to evaluate how much selected words are relevant in explaining the response variable when new similar datasets are considered. In fact, the datasets we analyse are renewed either at any season (in case of e-commerce data) or at any further session of Human Resources evaluation (in case of employees' ratings data) and the goal of our study is to identify which methods allows us to retrieve the relevant drivers of either item prices or employees satisfaction in any further dataset of this type. The analyses have been carried out for each case study. Subsequently we will compute the bootstrap inclusion frequency of selection for each variable and, when possible, we provide an assessment on the relevance of selected variables for the case study at hand. For each specification, we consider the set of features selected by each bootstrap replication and we sort them in decreasing order according to the number of times in which they have been selected. If some variables present a high number of repetitions it means that they have been selected in several bootstrap samples and thus that the robustness of model to small perturbations of the data set, i.e. the model validity, increases. As far as LDA is concerned, after discarding duplicates, the top terms of each topic in each bootstrap replication are considered. This allows us to preserve comparability. In this way, for each model, we evaluate the performance of alternative selection models on the basis of three criteria: 1) their explanatory power out of the training sample, measured in terms of Predictive R^2, i.e. in terms of explanatory power over new similar datasets; 2) the bootstrap inclusion frequency of selected variables; 3) an a priori evaluation of their truly influential values, when it may be formulated, like in the first case study where we may refer to Cachon and Swinney (2011). As an additional criterion, we provide summary statistics of the estimates of coefficients of the most frequently included variables over bootstrap replications: the lower the variability displayed by the estimates, the higher the robustness of the selection method. Here, only tables pertaining to 1) and 2) are shown but the overall analysis of the results includes considerations based on 3).

Several variants of lasso are considered. Firstly, to tune the lasso parameter, we recur to standard criteria used to optimise the predictive performance. We call lasso-min the model attained by minimising the cross-validation error, which is recognized to optimize predictive performance. Secondly, we optimise the tuning parameter by recurring to BIC variants: lasso-bic is attained by minimising the BIC, while lasso-ebic05 and lasso-ebic10 are attained by minimising the Extended BIC, with moderate and high model complexity regulated by the tuning parameter $\gamma = 0.5$ and $\gamma = 1.0$, respectively. To evaluate the performance of randomisation, we choose the Stability Selection method in the variant proposed by Shah et al. (2013), by imposing different thresholds of the selection

probability. The results are very similar by changing the selection probability and here we present those attained with the thresholds 0.7, named ssmb. As a screening method, we run the SIS algorithm and consider the performance produced by different number of selected predictors, form 1 to 25, labelled from sis-k1 to sis-k25. Finally, we evaluate the performance of the simple lasso obtained by imposing different number of selected predictors, from 1 to 25, labelled from lasso-k1 to lasso-k25.

We compare our results with those attained through the standard LDA analysis, which is the most popular topic models. We do not exploit the many variants of LDA for short text neither the supervised LDA as LDA is presented here only as a general benchmark, but the focus is on variable selection in regression models. Over the randomised training sets, we select LDA solutions corresponding to different number of topics and predict indicator variables for topics with probabilities larger than the mean topic probability in each micro sentence. Then, we regress the response variable, (either ratings o prices) over the indicator variables and evaluate the performance, according to the procedure previously exposed.

4. Results

4.1. Case study 1: Prices

The main results, in terms of number of selected variables, Adjusted and Predictive R^2, are displayed in Table 1, regarding Knitwear.

Table 1. Prices data set. Number of Selected Variables, Adjusted and Predictive R^2, and Inclusion Frequency. (Means and Standard Deviations over Bootstrap Replications)

Prices	Selected Vars		Adjusted R^2		Predictive R^2		Inclusion
Knitwear	mean	sd	mean	Freq.	mean	sd	Freq. mean
lasso min	53.6	23.7	0.828	0.056	0.773	0.095	0.612
lasso bic	24.7	6.6	0.802	0.067	0.698	0.101	0.610
lasso ebic0.5	19.1	6.1	0.775	0.079	0.661	0.108	0.597
lasso ebic1.0	14.2	6.1	0.732	0.101	0.613	0.120	0.570
lasso k=ebic1.0	14.0	0.9	0.732	0.059	0.624	0.107	0.614
sis k=ebic1.0	14.0	0.0	0.719	0.057	0.625	0.099	0.623
ssmb	6.3	1.5	0.580	0.086	0.499	0.114	0.620
lasso k=ssmb	6.0	0.7	0.594	0.068	0.493	0.115	0.554
sis k=ssmb	6.0	0.0	0.580	0.066	0.493	0.111	0.565
lda k=ssmb	6.0	0.0	0.116	0.063	0.158	0.073	0.429
lda k=ebic1.0	14.0	0.0	0.292	0.083	0.366	0.090	0.487

As expected, lasso-min always selects, on average, a very high number of predictors in both categories, to reach the highest Adjusted and Predictive R^2 and

by explaining 0.773 of price variation for knitwear replicated datasets. Indeed, as it has been observed it tends to be generous in selecting noisy variables. The number of predictors selected by lasso-min displays the highest standard deviation. When lasso is optimised by minimising the eBIC, it employs few parameters; the higher the value of the tuning parameter, the smaller the number of selected variables, on average. The most parsimonious eBIC selects, on average, 14.2 predictors, to explain from 73.2 % of the price variability in actual datasets and about 61% of the price variability in new datasets. Stability Selection is quite parsimonious as well: it selects in mean 6.3 predictors, explaining a share of 0.499 of prices' predictive variance within categories. Predictive R^2 are just lower than the ones of lasso-ebic10, though based on quite less predictors. Moreover, it has to be observed that lasso models based on the ebic optimization tend to select a limited number of coefficients on average, but with a certain degree of heterogeneity over the replications.

For sake of brevity, having results for k = 1,2,...,25, Table 1 presents results for SIS and lasso computed with the same fixed number of predictors as ssmb and lasso-ebic10, i.e. $P^* = 6$ (knitwear) and $P^* = 9$ (dresses) for Stability Selection and $P^* = 14$ (knitwear) and $P^* = 19$ (dresses) for lasso optimised with eBIC with $\gamma = 1$.

Focusing on the SIS method, we note that it produces very similar results as compared to ssmb and lasso-ebic10 when a comparable number of predictors are imposed. The same pattern may be observed when lasso is used by fixing the number of selected variables. As expected, the performance in term of Predictive R^2 increases, with diminishing derivative, as long as the number of predictors grows.

Concerning the comparison with topic models, we may observe that the results attained through text regression methods always significantly overcome the performance of LDA, in terms of Predictive R^2. Mean inclusion frequencies of LDA models are comparable to the ones attained by other methods. At the same time, the most relevant selected words are quite different from those selected by text regression and they more scantly contain information about materials.

4.1. Case study 2: Ratings

Table 2 displays the results of the second analysis, conducted with the aim of explaining ratings with tokens drawn by open questions. In this example Lasso-bic tend to select the largest number of variables, capable to explain around 78.4% of predictive ratings variation. As in the previous example, also in this case, lasso-min produces, on average, a high number of predictors, to reach the highest Predictive R^2. More parsimonious models are selected by ssmb and lasso-ebic10, based on less nine variables or less, even if also lasso optimizing the ebic yet displays higher variability for the number of coefficients. As in case study 1,

lasso and SIS perform comparably well as ssmb and lasso-ebic10, when the number of predictors is a priori fixed, with an out of sample R^2 oscillating around 30 to 35%.

Table 2. Ratings data set. Number of Selected Variables, Adjusted and Predictive R^2, and Inclusion Frequency. (Means and Standard Deviations over Bootstrap Replications)

Ratings	Selected Vars		Adjusted R^2		Predictive R^2		Inclusion
	mean	sd	mean	sd	mean	sd	Freq. mean
lasso min	171.0	37.4	0.829	0.049	0.851	0.067	0.488
lasso bic	213.5	159.1	0.806	0.191	0.784	0.223	0.390
lasso ebic0.5	16.8	32.9	0.427	0.090	0.390	0.087	0.520
lasso ebic1.0	7.0	17.7	0.325	0.065	0.302	0.067	0.651
lasso k=ebic1.0	7.0	0.6	0.334	0.030	0.319	0.042	0.745
sis k=ebic1.0	7.0	0.0	0.343	0.029	0.305	0.045	0.627
ssmb	8.8	1.6	0.369	0.040	0.338	0.049	0.670
lasso k=ssmb	9.0	0.6	0.362	0.030	0.345	0.043	0.724
sis k=ssmb	9.0	0.0	0.370	0.029	0.332	0.045	0.630
lda k=ssmb	9.0	0.0	0.343	0.044	0.390	0.051	0.528
lda k=ebic1.0	7.0	0.0	0.344	0.051	0.383	0.055	0.505

Differently than in the case of prices, in this circumstance, our findings are slightly in favour of topic modeling, as the LDA predictive R^2, corresponding respectively to solutions with 7 and 9 topics, to preserve comparability, are slightly higher than the ones reached through text regression methods. At the same time, LDA specifications are in relative delay in terms of inclusion frequency. In our view, the limited gain in predictive performance, whenever realized, does not compensate for the loss of robustness in selecting variables.

4. Concluding remarks

The paper has investigated the analytics that allow one to exploit the informative content and the explanatory power of unstructured, short texts on a response variable. Interpretability of the results was a key issue for the scope of the present study, hence we have restricted our focus to shrinkage methods and, within this class, we have favoured models that provide results of easier interpretation.

The results of the analysis of two case study attain insights into two main directions concerning, on one side, the performance of analysed models within the class of text regression, and, on the other side, the different ability of text regression versus topics modelling methods in extracting information from short texts. Along the first direction, our findings show that, in terms of explanatory power, both Stability Selection and Sure Independent Screening are able to

improve the lasso, when the latter is optimised through standard criteria finalised to prediction purposes. Nevertheless, by limiting the number of selected variables, lasso is capable to attain comparable results. As far as the ability to select relevant explanatory predictors is concerned, Stability Selection slightly outperformed the other methods. A relevant finding is that lasso behaves as well as alternative computationally more intensive methods when the number of selected variables is limited.

Concerning the comparison of text regression with LDA, the former outperforms LDA in terms of explanatory power in the prices case study whilst LDA outperforms text regressions in the ratings case study. This is likely to happen both because texts are longer in the ratings case study and because of their contents, which are naturally more connected to latent topics. In terms of the average repetition of the selected words, the opposite holds. Inclusion frequencies of LDA are higher than those of text regressions in the case of prices and lower in the case of reviews. However, words selected based on text regressions are always preferable to LDA ones, so that in both cases text regressions appear highly suitable to pick up relevant words within a bag of words. Our findings favour variable selection in text regressions as a method that may provide valuable solutions when texts are short.

References

Cachon G. P. and R. Swinney (2011). The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior. *Management Science*, 57(4), 778–795.

Efron B. and R. Tibshirani (1998). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Gentzkow M., Kelly B. and M. Taddy (2019). Text as data. *Journal of Economic Literature,* 57 (3), 535–574.

Hastie T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Being known, being known for something and being good: A new textual measure of organizational reputation

Andrea Fronzetti Colladon

University of Perugia – andrea.fronzetticolladon@unipg.it
Kozminski University

Abstract

Organizational reputation is a simple yet highly debated concept, with past studies presenting multiple definitions, conceptualizations and operationalization approaches (Lange et al., 2011). Building on past research, we extend the conceptualization of organizational reputation and present new methods and tools for its assessment from the analysis of (big) textual data. We bring together considerations from different research trends regarding semantic network analysis, sentiment analysis, topic modeling, natural language processing, and the study of honest signals in human communication. The combination of these methodologies opens up unprecedented opportunities to develop business analytics, make sense of big data, read the collective mind, and discover emergent trends and behaviors (Fronzetti Colladon et al., 2020).

Nowadays text data is ubiquitous and often freely accessible from multiple sources. From news and social media analysis, for example, we can understand the intentions, goals and values of corporate stakeholders (Barchiesi & Fronzetti Colladon, 2021a, 2021b) – considering more spontaneous communication than responses given to surveys. Indeed, we can look at honest signals, described as subtle patterns in how we interact and communicate (Pentland, 2008).

From stakeholder and media communications we assess organizational reputation, considering the dimensions of "being known", "being known for something" and "being good". In particular, we extend the conceptualization and applicability of the Semantic Brand Score (SBS) indicator (Fronzetti Colladon, 2018) to the measurement of organizational reputation. The Prevalence dimension of the SBS is used to capture the "being known" dimension, as visibility of a company name certainly favors its recognition and recall, increasing its awareness. The "being known for something" dimension relates to the construct of brand associations and the image that stakeholders have of a company. On the one side, having more associations is usually preferable, as they can increase brand memorability and knowledge (Keller, 1993). On the other side, it is also important to look at the quality and strategic value of these associations. Accordingly, we measure this second dimension of company reputation considering:

- the heterogeneity and uniqueness of brand associations, with the idea that more distinctive associations will help differentiate one company from the others (we use the distinctiveness centrality indicator) (Fronzetti Colladon & Naldi, 2020);

- the connective power of a company node in the semantic network (measured through the Connectivity dimension of the SBS);
- the strength of indirect associations, evaluated through a network approach to topic modeling (Fronzetti Colladon & Grippa, 2020).

Lastly, the "being good" dimension is analyzed but looking at the favorability of brand associations, through sentiment analysis.

Having a strong reputation usually leads to increased company performance and can influence, for example, brand-choice probability, managers and shareholders decisions, and market outcomes (e.g., Rovelli et al., 2022). We validate our approach with some case studies, also discussing the different impact of each dimension of reputation on business performance.

In this work, we put together large-scale data analytics with the consideration of honest signals in human communication. We present an extended conceptualization of organizational reputation and methods and tools for its measurement – with the ultimate goal of supporting managers, and other stakeholders, in their strategic decisions.

Keywords: organizational reputation; semantic brand score; semantic network analysis; big data; honest signals; text mining.

References

Barchiesi, M. A., & Fronzetti Colladon, A. (2021a). Corporate core values and social responsibility: What really matters to whom. *Technological Forecasting and Social Change*, *170*, 120907. https://doi.org/10.1016/j.techfore.2021.120907

Barchiesi, M. A., & Fronzetti Colladon, A. (2021b). Big data and big values: When companies need to rethink themselves. *Journal of Business Research*, *129*, 714–722. https://doi.org/10.1016/j.jbusres.2019.10.046

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

Fronzetti Colladon, A., & Naldi, M. (2020). Distinctiveness centrality in social networks. *PLOS ONE*, *15*(5), e0233276. https://doi.org/10.1371/journal.pone.0233276

Keller, K. L. (1993). Conceptualizing, Measuring, and Managing Customer-Based Brand Equity. *Journal of Marketing*, *57*(1), 1–22.

Lange, D., Lee, P. M., & Dai, Y. (2011). Organizational Reputation: A Review. *Journal of Management*, *37*(1), 153–184. https://doi.org/10.1177/0149206310390963

Pentland, A. (2008). *Honest Signals: How They Shape Our World*. MIT Press.

Rovelli, P., Benedetti, C., Fronzetti Colladon, A., & De Massis, A. (2022). As long as you talk about me: The importance of family firm brands and the contingent role of family-firm identity. *Journal of Business Research*, *139*, 692–700. https://doi.org/10.1016/j.jbusres.2021.09.075

Text mining e analisi ermeneutica: il caso delle pratiche di volontariato internazionale

Augusto Gamuzza[1], Francesca Greco[2], Anna Maria Leonora[1]

[1] University of Catania – augusto.gamuzza@unict.it
anna.leonora@unict.it
[2] University of Rome La Sapienza – francesca.greco@uniroma1.it

Abstract
It is well known that the biographical interview is a tool used to gather information about the interviewee's experience and worldview. There are different techniques for analysing interviews, both qualitative and quantitative. This paper describes a methodology for the analysis of interviews, based on a mixed-method approach, which integrates statistical techniques of textual analysis with hermeneutic analysis in order to facilitate the work of the researcher and to offer a synthetic, structured and in-depth view of the information collected. The procedure includes a first phase in which the method of Emotional text Mining (ETM) is used to analyse the interviews with the aim of constructing a synthetic framework that returns the symbolic-cultural matrix and identity representations emerging from the discourses of the interviewees. This procedure provides an outline that supports the subsequent phase of hermeneutic content analysis, which allows us to observe the more specific aspects of the narratives placed in connection with the theoretical model that guides the research. The procedure proceeds from a maximum level of synthesis, which is more easily reached through statistical analysis, to a maximum level of detail of the information, thanks to the use of the interpretative tools of "classical" qualitative analysis. The method described does not only provide a guide that facilitates the work of the researcher, but provides an overall framework that supports the process of understanding the phenomenon under investigation.

Keywords: Text-mining, Hermeneutic Analysis, Cosmopolitanism, Solidarity practices, Volunteering, NGO

Riassunto
È noto che l'intervista biografica è uno strumento utilizzato per raccogliere informazioni sull'esperienza e la visione del mondo dell'intervistato. Ci sono diverse tecniche per analizzare le interviste, sia qualitative che quantitative. Questo articolo descrive una metodologia per l'analisi delle interviste, basata su un approccio misto, che integra tecniche statistiche di analisi testuale con l'analisi ermeneutica al fine di facilitare il lavoro del ricercatore e di offrire una visione sintetica, strutturata e approfondita delle informazioni raccolte. La procedura prevede una prima fase in cui il metodo dell'Emotional text Mining (ETM) viene utilizzato per analizzare le interviste con

l'obiettivo di costruire un quadro sintetico che restituisca la matrice simbolico-culturale e le rappresentazioni identitarie emergenti dai discorsi degli intervistati. Questa procedura fornisce uno schema che supporta la successiva fase di analisi ermeneutica del contenuto, che permette di osservare gli aspetti più specifici delle narrazioni poste in connessione con il modello teorico che guida la ricerca. La procedura procede da un livello massimo di sintesi, più facilmente raggiungibile attraverso l'analisi statistica, ad un livello massimo di dettaglio delle informazioni, grazie all'utilizzo degli strumenti interpretativi dell'analisi qualitativa "classica". Il metodo descritto non fornisce solo una guida che facilita il lavoro del ricercatore, ma fornisce un quadro generale che sostiene il processo di comprensione del fenomeno indagato.

Parole chiave: Text-mining, analisi ermeneutica, cosmopolitismo, pratiche di solidarietà, volontariato, ONG

1. Introduzione

L'intervista biografica come strumento di ricerca utilizzato per raccogliere informazioni sull'esperienza e la visione del mondo dell'individuo richiede spesso al ricercatore una sensibilità analitica ulteriore che riconduca l'evento idiografico (la stora di vita di un singolo soggetto) ad un livello di sintesi più elevato che includa la dimensione macro (e storica) del mutamento sociale (Bertaux, 1981). La proposta di questo lavoro è presentare in modo argomentato quale sia il valore aggiunto del trattamento statistico dei dati testuali tramite *Emotional Text Mining* (ETM) (Greco, 2016; Greco and Polli, 2020) come propedeutico ad un'analisi tematica del contenuto. ETM attraverso un approccio socio-costruttivista con orientamento psicodinamico considera l'intervista biografica un prodotto delle simbolizzazioni culturali condivise dai soggetti coinvolti nell'evento comunicativo (intervista). In questa prospettiva, il ricercatore e il soggetto intervistato co-elaborano le categorie logiche del proprio discorso (stimolo-risposta narrativa) determinando le loro interazioni, aspettative e scelte comunicative: l'approccio statistico consente di estrarre gli aspetti simbolico culturali le rappresentazioni sociali connesse con la scelta di agire all'interno di specifici contesti. La ricerca da cui è tratto il corpus si colloca nel quadro di riferimento teorico più generale dei *Global Studies* con l'obiettivo di indagare la dimensione solidale della socializzazione cosmopolita (Cicchelli, 2018) osservando una particolare tipologia di soggetti: i volontari per lo sviluppo nei paesi terzi che operano all'estero attraverso ONG Italiane. Più specificamente, la ricerca ha indagato come le pratiche di solidarietà transnazionale impattano sui percorsi biografici evidenziandone legami e contraddizioni con le pratiche di solidarietà cosmopolita. Per affrontare queste sfide, è stato scelto un modello analitico (Cicchelli, 2018) che incorpora la dimensione processuale degli apprendimenti cosmopoliti – la socializzazione cosmopolita - definendola come un processo riflessivo attraverso il quale gli

individui sviluppano e alimentano il rapporto con l'alterità. Questa proposta offre una prospettiva concreta per la formulazione di strumenti che catturino la dimensione empirica delle pratiche di solidarietà messe in atto dai volontari per lo sviluppo . Secondo la proposta epistemologica che fa da spina dorsale a questo lavoro, al fine di inquadrare la processualità sottesa all'apprendimento della relazione individuale con l'alterità sono possibili quattro forme elementari cosmopolite che definiscono l'approccio dell'individuo alla realtà sociale globale: cosmo-estetica, cosmo-culturale, cosmo-etica e cosmo-politica (Cicchelli, 2018). In particolare, per i nostri scopi euristici, abbiamo concentrato la nostra attenzione sulla dimensione cosmo-etica dello spirito cosmopolita, individuandone una traduzione operativa. Lo spirito cosmo-etico è in linea con gli elementi costitutivi della solidarietà cosmopolita: questo orientamento presuppone che la fedeltà a uno Stato-nazione non impedisca il riconoscimento del sé nella comune umanità che è, in linea di principio, il fondamento stesso dell'estensione della solidarietà oltre i confini nazionali (Cicchelli, 2018). Pertanto, il focus sull'etica cosmopolita riflette le traiettorie biografiche dei soggetti la cui caratteristica principale è la cura per gli altri distanti (*distant others*): i vololontari sono assumibili ad attori sociali che non possono 'sfuggire' al dovere morale di agire attraverso l'azione di cooperazione.
In maniera sintetica, questo lavoro cercherà di rispondere alle seguenti domande: qual è il valore aggiunto in termini analitici nell'integrazione di ETM con tecniche di analisi del contenuto rispetto ad un corpus medio grande di interviste biografiche? Quali punti di forza e di debolezza si possono individuare in questo percorso analitico? Come scegliere la sequenza di integrazione?

2. Metodologia

Sebbene il più generale impianto della ricerca e lo stesso oggetto d'indagine abbiano imposto di scegliere uno strumento di raccolta dati di tipo qualitativo, *insider*, questo articolo approfondisce criticamente la metodica applicata per l'analisi dei dati testuali raccolti. Tale strategia analitica, che ha integrato una tecnica statistica di analisi testuale (*ETM*) con la più classica analisi ermeneutica del contenuto biografico, è volta a costruire una visione più estesa delle informazioni raccolte, nonché formulare un'interpretazione che integri al tempo stesso un livello sintetico (tecniche quantitative) e un elevato dettaglio (tecniche qualitative).
Tale percorso è stato posto in essere attraverso la realizzazione di interviste biografiche (29) che sono state condotte da un team di rilevatori opportunamente formati fra il 2017 ed il 2021. Lo studio è stato condotto all'interno di un gruppo di 5 ONG italiane operanti in Africa, Sud America e Medio Oriente con un campionamento ragionato che ha coinvolto 29 soggetti in base alle loro posizioni istituzionali considerando un'ampia estensione dei ruoli legati alla cooperazione

internazionale (volontari; manager dei volontari; project manager, country manager, esperti tecnici, operatori sociali, servizio civile all'estero). Gli intervistati sono prevalentemente donne (22 su 29) con un'età media di 36 anni (SD 4,7) e con un alto livello d'istruzione: 18 avevano ottenuto una laurea (un intervistato frequenta a un corso di dottorato), 5 un master, mentre uno è attualmente studente universitario. La durata dell'esperienza con le ONG internazionali varia da un'esperienza brevissima (fino a 1 anno=31%), medio-breve (da più di 1 anno a 3 anni = 31%), a un'esperienza di medio periodo (da 4 a 10 anni = 21%) e lunga (più di 10 anni = 17%).

La procedura applicata all'analisi prevede una prima fase in cui la tecnica ETM (Greco and Polli, 2020) viene utilizzata per trattare il corpus delle interviste con l'obiettivo di costruire un quadro sintetico che restituisca la matrice simbolico-culturale e le rappresentazioni identitarie emergenti dai discorsi degli intervistati. Questa procedura fornisce uno schema che supporta la successiva fase di analisi ermeneutica del contenuto, che permette di osservare gli aspetti più specifici delle narrazioni poste in connessione con il modello teorico che guida la ricerca. Pertanto, l'analisi si è svolta da un livello massimo di sintesi, più facilmente raggiungibile attraverso l'analisi statistica, ad un livello massimo di dettaglio delle informazioni, grazie all'utilizzo degli strumenti interpretativi dell'analisi qualitativa "classica". Il metodo descritto non fornisce solo una guida che facilita il lavoro del ricercatore, ma fornisce un quadro generale che sostiene il processo di comprensione del fenomeno indagato.

A questo scopo le trascrizioni delle interviste sono state unite in un corpus analizzato con il metodo dell'ETM. Inizialmente i testi sono stati pretrattati e sono stati calcolati tre indicatori lessicali per valutare la possibilità di processare statisticamente i dati: i token, il type-token ratio e la percentuale di hapax (Giuliano and La Rocca, 2010). Successivamente, si è effettuata la suddivisione del corpus in unità di contesto, la lemmatizzazione, e sono state escluse dall'analisi le parole vuote, i termini di rango alto di frequenza e quelli del rango basso di frequenza fino ad un valore pari al numero di interviste al fine di ridurre la dimensione della matrice (Greco, 2016). Sulla matrice termini-unità di contesto è stata effettuata un'analisi dei cluster con un algoritmo bisecting *k*-mean, limitato a 20 partizioni, con l'esclusione delle unità di contesto che non presentavano almeno una co-occorrenza, ed è stato utilizzato il coseno come misura di similarità (Steinbach et al., 2000). Il Calinski-Harabasz, il Davies-Bouldin e il coefficiente di correlazione intraclasse sono stati calcolati per identificare la partizione ottimale e sulla matrice termini-cluster è stata effettuata una analisi delle corrispondenze (Lebart e Salem, 1994) per identificare gli assetti simbolici caratterizzanti la comunicazione (Cordella et al., 2014; Greco, 2016). Al fine di selezionare alcune unità di contesto rappresentative delle tematiche individuate è stato calcolato lo *score*, vale a dire la somma dei valori pesati del

χ^2 dei termini che co-occorrono in un'unità di contesto classificata in uno specifico cluster.

3. Risultati

L'analisi quantitativa ETM sul corpus delle interviste è stata implementata prima dell'analisi ermeneutica qualitativa al fine di perimetrare un set di temi emergenti dalle narrazioni contenute nelle interviste raccolte. I 391 termini selezionati hanno consentito di classificare l'89% delle interviste e sono stati identificati 8 temi principali (cluster) e l'analisi delle corrispondenze ha identificato 7 dimensioni latenti (fattori che organizzano le narrazioni) di cui le prime cinque spiegano il 78% dell'inerzia. In tabella 2 è riportato il quadro sintetico dell'interpretazione dello spazio fattoriale e dei cluster in ragione delle parole che li caratterizzano.

Tabella 1. Quadro sintetico dell'interpretazione dell'analisi multivariata delle interviste (tra parentesi accanto al fattore è riportata la percentuale di inerzia spiegata e sotto l'etichetta dell'asse fattoriale il valore della coordinata del centroide del cluster)

Cl	*UC%*	*Temi*	*Fattore 1 22,0%*	*Fattore 2 16,2%*	*Fattore 3 14,9%*	*Fattore 4 12,9%*	*Fattore 5 12,0%*
			Motivazione	*Contesto*	*Viaggio*	*Posizionam. Identitario*	*Esperienza*
1	11,3	Esplorazione esistenziale	Personale 0,32	Scoprire -0,30	-0,12	0,06	Ricordi -0,36
2	11,6	Esperienza di cooperazione	Istituzionale -0,51	Scoprire -0,25	Ritorno 0,37	Paritario -0,47	Ricordi -0,22
3	14,0	Turismo vocazionale	Personale 0,36	0,07	-0,03	Paritario -0,55	Emozioni 0,45
4	13,5	Professionismo coloniale	Istituzionale -0,57	Cambiare 0,26	Partenza -0,52	Asimmetrico 0,19	Emozioni 0,30
5	14,6	Modalità operativa	Personale 0,39	Cambiare 0,70	Ritorno 0,28	0,15	Ricordi -0,19
6	15,6	Programmazione della partenza	Personale 0,44	Scoprire -0,47	Partenza -0,19	Asimmetrico 0,26	0,06
7	9,4	Arrivo in loco	Istituzionale -0,40	0,09	Partenza -0,36	-0,11	Ricordi -0,56
8	10,1	Riadattamento	Istituzionale	Scoprire	Ritorno	Asimmetrico	Emozioni
			-0,42	-0,21	0,69	0,47	0,31

UC = unità di contesto

Nel complesso, gli intervistati lavorano nel campo della cooperazione internazionale attraverso le ONG sulla base di una motivazione personale o perché già inseriti in specifici percorsi istituzionali di professionalizzazione. Sono motivati dal desiderio di attivare un cambiamento nella propria vita o fare nuove esperienze lontano da casa e le interviste si organizzano su due momenti distinti del viaggio come esperienza di vita: la partenza ed il ritorno in Italia. Tale scansione temporale incide sul processo di posizionamento rispetto ai beneficiari

degli interventi di cooperazione (i cosiddetti "locali") che viene vissuto come paritario o asimmetrico, in un'ottica che in alcuni casi sembra scivolare verso un atteggiamento "colonialista". La propria esperienza, infine, viene narrata attraverso le emozioni, a volte particolarmente intense, o i ricordi.
Entro questi assi della comunicazione si collocano i diversi modi di rappresentare e declinare la propria esperienza come: un'esplorazione esistenziale, un'esperienza di cooperazione in senso tecnico-professionale, una forma di turismo "vocazionale", una pratica di "professionismo coloniale". Inoltre, si sono individuati ulteriori contesti narrativi rispetto alla rappresentazione della propria esperienza: modalità operative, programmazione della partenza, il processo di inserimento a seguito dell'arrivo nel luogo di intervento e il riadattamento necessitato dalla differenza fra i diversi contesti di vita (Africa-Italia). La successiva analisi ermeneutica si è inserita e integrata alle prime evidenze così ottenute. Secondo la più classica delle prospettive biografiche (Ferrarotti, 1999; Rosenthal, 1993) l'analisi sul narrato biografico restituisce l'insieme significativo (per l'attore sociale) delle interazioni fra soggetto e contesto sociale situazionale. In particolare, per i nostri scopi euristici, ci si è riferiti alla dimensione cosmo-etica dello spirito cosmopolita, (Cicchelli 2018) come categorie nodali per il processo interpretativo delle traiettorie biografiche dei soggetti la cui caratteristica principale è la cura per gli altri lontani, un attore sociale che non può 'sfuggire' al dovere morale di agire.
L'analisi ermeneutica, è stata condotta da una prospettiva socio-figurazionale (Elias ; Rosenthal 2019) a partire dall'individuazione di alcune categorie iscrivibili nella più ampia cornice teorica dei *global studies* (tabella 1).

Tabella 2. Le categorie interpretative per l'analisi ermeneutica delle interviste biografiche

Assi analitici	Livelli di osservazione	Dimensioni analitiche
Lezioni Apprese	Macro	Ambiente relazionale cosmopolita
	Meso	Ambiente relazionale strutturato dai ruoli nell'organizzazione
	Micro	Ambiente relazionale individuale
MOTIVATION	Macro	Riduzione della disuguaglianza
	Meso	Relazione di aiuto e cura
	Micro	Empatia e rapporto con la diversità
SOLIDARITY AS ETHICAL PRINCIPLE	Macro	Registro spaziale
	Meso	Registro temporale
	Micro	Fonti di solidarietà e meccanismo di apprendimento
SKILLS AND TOOLS	Macro	Competenze e strumenti come visione globale
	Meso	Competenze e strumenti della struttura associativa
	Micro	Competenze individuali di base
SOLIDARITY AS ACTION	Macro	Il volontariato come cittadinanza universale
	Meso	Relazione con l'organizzazione NGO
	Micro	Cooperazione come interazione della vita quotidiana

4. Discussione e conclusioni

Al fine di rispondere alle ambiziose domande che abbiamo posto all'inizio appare utile, dato il carattere esplorativo della ricerca, suddividere queste brevi note conclusive lungo due direttrici argomentative: una metodologica e una teorica. In prima istanza, considerando la complessità sottesa all'operativizzazione di concetti e processi tanto "densi"; si sono individuati, attraverso la tecnica ETM un *set* di temi emergenti rispetto all'esplorazione dei differenti percorsi biografici sottoposti ad analisi. Nel nostro caso, con un corpus di dimensioni medio grandi, la tecnica ETM è servita a ricostruire le interrelazioni narrative tra gli eventi che i soggetti vivono, le loro esperienze, le loro memorie e il modo attraverso cui ne parlano. La ricerca ha quindi puntato ad individuare i meccanismi che stanno 'dietro' il racconto dell'esperienza; i contesti di costruzione narrativa e come le persone le rappresentano: come parlano, nel presente dell'atto di intervista, della propria esperienza strutturata nel passato ma che continua ad evolversi. In questo quadro, il rapporto tra teoria e ricerca si è configurato in prospettiva abduttiva: la risorsa della narrazione biografica diviene metafora di una conoscenza profonda del vissuto esperienziale e valoriale del soggetto; essa è letta e interpretata dalla lente teorica consolidata del concetto di *solidarietà cosmopolita* che permette di esprimere e osservare non solo tipologie esperienziali ma anche dinamiche relazionali strutturalmente situate (micro-meso-macro). L'opzione di applicare un *mixed textual method*, offre dunque al ricercatore dei livelli di comprensione ulteriore rispetto al processo di ermeneutica 'classica' del dato narrativo realizzando un processo di sintesi delle informazioni raccolte non praticabile se non attraverso un'analisi computer-assistita che assicura una gestione controllata di un corpus di interviste come quelle del presente caso.

Dal punto di vista più strettamente teorico, l'analisi svolta ha permesso di mettere a fuoco in modo innovativo le dinamiche che traducono l'esperienza del volontariato internazionale in elementi di significato biografico, consentendo - grazie al dialogo analitico stabilito fra le due tecniche di analisi - un approfondimento ragionato dei dati narrativo-testuali e la possibilità di meglio perimetrare, grazie a una lente analitica integrata, l'insieme degli elementi contraddittori della dimensione identitaria che interagiscono con la creazione di repertori di conoscenza post occidentali e non sempre cosmopoliti.

Inoltre, la ricerca ha evidenziato anche un'ulteriore spazio di riflessione in ordine al momento (nel senso temporale) in cui trattare il testo tramite le tecniche quantitative di text mining. Nel nostro caso, ETM è stata realizzata prima dell'analisi ermeneutica ma sarebbe altrettanto interessante vedere quali elementi interpretativi potrebbero provenire da una inversione nell'applicazione delle due tecniche, qualitativa e quantitativa nel trattamento del testo proveniente da interviste biografiche. In questo senso, si apre una direttrice di ricerca promettente che richiede ulteriori, necessari, approfondimenti.

Bibliografia

Breckner R. e Massari M. (2019). Past, present and future of biographical research. A dialogue with Gabriele Rosenthal. *Rassegna Italiana di Sociologia*, 1: pp. 155-184

Cordella B., Greco F. e 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.

Cicchelli V. (2018). *Plurale e comune. Sociologia di un mondo cosmopolita*. Morlacchi.

Ferrarotti F. (1999). *L'ultima Lezione. Critica della sociologia contemporanea.* Laterza.

Giuliano L. e 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. e Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management,* 51: 101934.

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

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

Rehearsals of ideological placement on a corpus of parliamentary questions

Franco M. T. Gatti

Università degli Studi di Padova - franco.gatti.1@phd.unipd.it

Abstract
The aim of this work was to analyse a corpus made of 96 Italian parliamentary questions from Movimento 5 Stelle, 96 from Partito Democratico (i.e. PD, the main Italian centre-left party), and 96 from Popolo delle Libertà (i.e. PDL, the Italian centre-right coalition at that time), for a total of 288 questions extracted from the XVII legislature shorthand reports (2013-2017), in order to see whether there can be found some linguistic remarks of the proximity of M5S to PD or PDL. Machine learning analyses conducted through the classification algorithm known as Support-Vector Machine (SVM) showed that M5S questions seem to be closer to the left party (PD) rather than to the right coalition (PDL). Furthermore, deeper stylometric analyses conducted through text clustering showed that M5S deputies result as perfectly merged within the political and institutional system of Italy, with more tendences towards the left (PD), confirming the results of the machine learning classification.

Keywords: machine learning, classification algorithms, parliamentary questions, Italian linguistics.

1. Introduction

The Movimento 5 Stelle (M5S) is an Italian populist party founded in 2009 by Beppe Grillo (a well-known Italian comedian) and Gianroberto Casaleggio, following the success of two years of rallies in which the then political class was mocked and criticised by the comedian. Since its foundation, the ideological alignment of M5S among Italian political parties has been subject of debate in public opinion. While its members always declared themselves as not politically aligned nor with the left traditional ideologies, neither with the right ones, many political scientists, journalists and analysts considered the party as a mixture of ideas taken from left and right parties. Political ideas are expressed by means of words, and texts make an evidence of what a party is and what it does spread. Anyway, differences occur between propaganda texts (e.g. websites, election posters) and texts that allow the life cycle of the legislatures within the most important institution of the Country. And differences occur, as well, between the genre of the political speech inside or outside the Parliament (Giuliano and

Villani, 2015), and other kind of texts that do not primarily carry persuasive purposes, but they rather have informative ones. Parliamentary questions consist in a text genre that in Italy presented some evolutions throughout the years and nowadays has taken a generally standardized structure, and within this structure deputies have room for few personal stylistic traits.
The aim of this work is to analyse a corpus of parliamentary questions from M5S, PD and PDL in order to see whether a machine learning algorithm for classification purposes and a text clustering analysis tend to attribute the M5S questions closer to PD, to PDL or to none of them. The Partito Democratico (PD) is still nowadays the centre-left party in Italy, while the Popolo della Libertà (PDL), which was the centre-right coalition, divided into parties with different rightist souls. Moreover, this work is part of a larger project that aims to analyse the evolution of Italian language within the textual genre of parliamentary questions extracted from the Italian Parliament, on a more substantial corpus of 3 millions occurrences circa, ranging from 1980 to 2021.

2. Corpus

The corpus analysed in this work consists in a total of 288 parliamentary questions extracted from the XVII legislature shorthand reports: 96 parliamentary questions from each PD, PDL and M5S were taken into account, with a pace of 20 questions for each year of the legislature, for each party. The only exception has been 2013 - the starting year of the legislature - which featured only 16 questions per party: this is due to the fact that at the beginning of every legislature it takes some time to bring back the regular pace in the routine works of the Parliament, especially for non-legislative activities. The selection of the parliamentary questions in this corpus was made regardless of a focus on specific topics. Since the objective is to analyse language and style, by selecting questions on the same topics, which is actually one of the parameters of research in the Parliament database, the risk is to get biased analyses, influenced by the presence among high and medium frequencies of words that are specific to a topic, but not meaningful for a study that aims to capture and analyse purely linguistic aspects.
Through the software TaLTaC 2 (Bolasco, 2010), the corpus was organised, tokenized and some preliminary measures were calculated. After the tokenization of the corpus, the total amount of word tokens resulted in 209,309 words, while the total amount of word types consisted in 17,995. A types/tokens ratio of 8,60%, together with a hapax percentage of 44,39% certified the tractability of the corpus with quantitative analysis methods as reliable (Bolasco, 2010). With the following distribution, the total amount of occurrences results as enough balanced to be representative of each party: 74,668 for M5S, 66,427 for PD, and 68,214 for PDL. The content words with the highest overall frequency within the corpus are *Ministro* (Minister), *legge* (law), *articolo* (article), *euro*, *Governo*

(government), *decreto* (decree), *sapere* (to know), *premesso* (given that), *Italia*, *Ministero* (Ministry). While many of these words refer to the ritual formula used to open Italian parliamentary questions from the 1980s onwards, the words *law*, *decree* and *article* indicate that a large proportion of the parliamentary questions in this corpus are asking about legislative measures. This fact is not to be taken for granted if one considers that the parliamentary questions as a genre cover a very wide range of topics, from specific cases concerning local facts to topics of international importance. Further and deeper qualitative and quantitative analyses concentrated not just on the analysis of individual words, but also of more linguistic features as character bigrams, character trigrams and word bigrams (i.e. sequences of two words).

3. Methods

Support-Vector Machines (SVM) is a well known group of state-of-the-art supervised machine learning classification algorithms, frequently used in many research fields, and among them in the textual data analysis one (Joachims, 1998; Mikros and Perifanos, 2013). The concept beneath SVM is to build in a geometrical space the ideal hyperplane outlined by the major distance between the closer points of two different classes. This hyperplane serves as a classifier, able to distinguish among data points of the two classes with the best accuracy among all the possible hyperplanes that could be built between them. Through the so called kernel trick, using a polynomial kernel, we're able to extend this concept through more than 2 dimensions, bringing it to as many dimensions as the number of classes we want to use for the classification task. In the case of this work, we are working on 3 classes (M5S, PD, and PDL).

The whole corpus of 288 parliamentary questions was divided into training set and test set. In order to see whether M5S questions are more similar to PD or PDL ones, the 192 questions from PD and PDL were assigned as training set, while the 96 from M5S have been isolated as a test set. Through the AMNP procedure (Mikros and Perifanos, 2013) running on R (R Core Team, 2013), the whole corpus was split in text chunks of 1,000 words each with a normal sampling, and some useful linguistic features have been taken into account from them. The AMNP procedure is structured in such a way as to sample the corpus in text chunks on which to perform the analyses: it is through the assignment of each text chunk to its respective class that the machine learning procedure provides the classification results. As for the linguistic features, more specifically, the 500 most frequent character bigrams, the 500 most frequent character trigrams, the 500 most frequent words and the 500 most frequent word bigrams were extracted from the text chunks and put in a matrix with as many rows as the number of text chunks of the training set, and as many columns as the number of linguistic features extracted this way. Since the 96 parliamentary questions

from PD were split in 64 text chunks and the 96 from PDL were split in 66, the resulting matrix had 130 rows, while it had 2,000 columns. The value inside a Z (X, Y) record of the matrix consists in the relative frequency of the linguistic feature Y within the text chunk X. Since some values presented a variance close to 0, the linguistic features carrying those values were removed in order to reduce the noise within the classification process. The final dataset, cleaned of the low variance variables then consisted in a matrix with 130 rows and 1,785 columns. At this point, after training the algorithm on the PD+PDL set, a 10 fold cross-validation was performed and, from it, a confusion matrix was produced, giving an accuracy of 79%. With the training done, the subsequent task consisted in testing the M5S subcorpus, by classifying each one of the 72 text chunks resulting from it in the categories "PD" or "PDL" outlined accordingly to the linguistic features employed in the previous training operations.
A second step of the research was to make a cluster analysis on the same corpus (Eder, 2017). This result was achieved through the R package Stylo (Eder, Rybicki, Kestemont, 2016), and it took into account the same text chunks previously analised through the AMNP (i.e. chunks of 1000 words each) obtained with the normal sampling. The linguistic features taken into account in this cluster analysis are the character trigrams, and a *Classic Delta* distance was applied as analysis algorithm.

4. Results

The confusion matrix obtained after training the SVM algorithm on the subcorpus of 192 parliamentary questions from PD and PDL shows that the accuracy in the classification task reaches the 79%. Even if it doesn't break the 80% threshold, this value can be considered good enough for reliable classifications, and the Precision, Recall and F1 Score parameters corroborate it as well.

Table 1. Confusion matrix obtained with the cross-validation procedure after the SVM training

SVM
Accuracy**: 0.7923**

	Reference	
Prediction	PD	PDL
PD	**5.5**	**1.8**
PDL	**0.9**	**4.8**

Precision (PD / PDL): **0.75 / 0.84**
Recall (PD / PDL): **0.86 / 0.73**
F1 Score (PD / PDL): **0.80 / 0.78**

The following task consisted in the actual attribution of M5S text chunks to PD or PDL classes, defined as in the training process. This operation resulted in **61** out of **72** text chunks attributed to PD, and the remaining **11** classified as PDL, as it is shown in the scheme that follows, which represents each M5S' text chunk attribution:

[1] PD PD PD PD PD PD PD PD PDL PD PD PDL PDL PD PD PD PD PD PD PD
[21] PD PD PD PD PD PD PD PD PD PDL PD PD PDL PD PDL PDL PD PD PDL PD
[41] PD PD PD PD PD PD PD PD PD PD PD PD PDL PD PD PD PD PD PD PD
[61] PD PD PD PD PD PD PDL PD PDL PD PD PD

It is interesting to observe that, in most cases, M5S text chunks that are classified as PDL are not really distant each others, testifying that there might be specific parliamentary questions that take on the stylistic features proper of the PDL class. In order to deepen the stylometric analysis of these sets of parliamentary questions, a cluster analysis through *Classic Delta* distance on the character trigrams extracted from the corpus was performed. A culling procedure was applied setting the minimum value at 20 and the maximum at 60. As Eder *et alii* (2016) specify, this procedure aims at excluding from the analyses a certain group of words: in this case, only words that are featured in at least the 20% of texts and not over the 60% of texts are considered in the analysis, while all the others are discarded. In other words, the 40% most frequent words in the corpus (a great part of which are function words) and and the 20% least frequent words (which can be considered as specific to certain topics covered in the parliamentary questions) are discarded, and only the others contribute to the analyses. Picture 1 presents the dendrogram resulting from this cluster analysis.

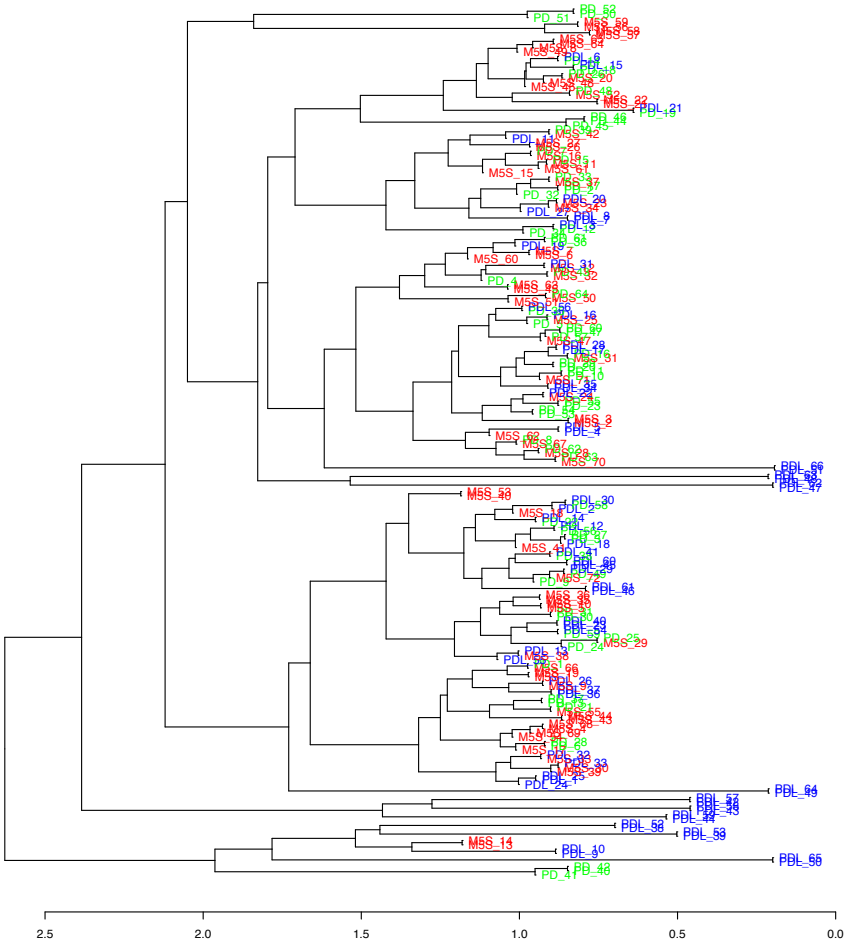


Figure 1. Cluster analysis on character trigrams, with a Classic Delta, culled from 20% to 60%.

By looking at the results of the cluster analysis it is possible to observe that, while some PDL clusters (color blue) are identifiable as in their own, M5S (color red) and PD (color green) tend more frequently to merge, outlining more similar stylistic traits.

5. Discussion and conclusions

From both the machine learning analysis and text clustering, two things seem to be clear:

1- The stylistic traits that the deputies of M5S expressed in their parliamentary questions from 2013 to 2017 are closer to the center-left party (PD) rather than

the center-right coalition (PDL). To those who are familiar with Italian politics, this may sound as no surprise, since in 2013 a post-electoral alliance was proposed from the PD leader at that time, Pier Luigi Bersani, to the M5S leaders, who declined his proposal. Moreover, since September 2019, M5S and PD have been together in the same government majority in two different Governments (Giuseppe Conte II and Mario Draghi);

2- the standardized structure of a parliamentary question certainly poses some limits in terms of freedom of expressiveness for a deputy, in this particular type of document. However, even with reduced stylistic possibilities to express their personal style, deputies tend to develop common stylistic features based on their political affiliation. These stylistic traits gave a direction in the classification with SVM, and later emerged clearly in the cluster analysis on the character trigrams, showing that M5S deputies are not just fully merged into the institution of the Italian Parliament, but they also tend to have more affinities to the left side of the political affairs within the very same institution.

In conclusion, from the perspective of a textual data analysis on the parliamentary questions, it seems to be clear that M5S members, who started their political journey into the Italian institutions declaring that they would shake up the Italian Parliament from every possible point of view, including an improvement in the quality of its outcomes and documents that regulate its daily life, had to adapt to the patterns and styles that have settled over the years, finding their own space mainly closer to the center-left side of the politics.

References

Bolasco S. (2010). *TaLTaC 2.10: sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi.* Milano: Led.

Eder M., Rybicki J. and Kestemont M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8 (1), 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Eder M. (2017). Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, 32(1): 50-64.

Giuliano L and Villani P. (2015). *Il linguaggio della leadership politica tra la Prima e la Seconda Repubblica, problemi di metodo e linee di ricerca*. Roma: Camera dei Deputati.

Joachims T. (1998). Text Categorization with Support Vector Machines. Proc. European Conf. Machine Learning (ECML'98). 10.17877/DE290R-5097.

Juola P., (2008). Authorship attribution, Foundations and Trends. *Information Retrieval* 1(3), 233–334.

Koppel M. and Seidman S. (2018). Detecting pseudoepigraphic texts using novel similarity measures. *Digital Scholarship in the Humanities*, 33(1), 72–81.

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.
Ondelli S. (2020). Semplicità, semplificazione, semplicismo. Su alcuni fraintendimenti nell'analisi della lingua dei politici. *Quaderns d'Italià*, 25, pp. 171-188.
R Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL: http://www.R-project.org/.
Savoy J. (2020). *Machine Learning Methods for Stylometry. Authorship attribution and author profiling.* Cham (Switzerland): Springer.
Tuzzi A. (Ed.) (2018). *Tracing the life cycle of ideas in humanities and social sciences*. Switzerland: Springer.

Finito di stampare nel mese di luglio 2022
per conto delle Edizioni Erranti
da Universalbook srl
Rende - Cosenza