

# Jadt 2010

## Statistical Analysis of Textual Data

Proceedings of 10th International Conference

*Journées d'Analyse statistique des Données Textuelles*

9-11 June 2010 - Sapienza University of Rome

Sergio Bolasco

Isabella Chiari

Luca Giuliano

editors

**JADT 2010**<http://jadt2010.uniroma1.it> <http://jadt.org>**Programme Committee**

Ludovic Lebart	CNRS - ENST, Paris ( <i>JADT President</i> )	Luca Giuliano	Sapienza Univ. di Roma
Ramon Alvarez-Esteban	Univ. de León	Benoît Habert	ENS-LSH Univ. Lyon 2
Enrica Aureli	Sapienza Univ. di Roma	Serge Heiden	ENS-LSH - ICAR Univ. Lyon 2
Harald Baayen	Univ. of Alberta	Alain Lelu	Univ. de Franche-Comté
Simona Balbi	Univ. di Napoli 'Federico II'	Damon Mayaffre	Univ. Nice Sophia-Antipolis
Valérie Beaudouin	Telecom ParisTech	Bruno Mazzara	Sapienza Univ. di Roma
Mónica Bécue	Univ. Politècnica de Catalunya	Sylvie Mellet	Univ. Nice Sophia-Antipolis
Sergio Bolasco	Sapienza Univ. di Roma	Denis Monière	Univ. de Montréal
Etienne Brunet	Univ. Nice Sophia-Antipolis	Bénédicte Pincemin	ICAR Univ. Lyon 2
Lou Burnard	Univ. of Oxford	Gérald Purnelle	LASLA Univ. de Liège
Isabella Chiari	Sapienza Univ. di Roma	Martin Rajman	EPFL Univ. de Lausanne
Claude Condé	Univ. de Franche-Comté	Max Reinert	CNRS / Univ. de Versailles
François Daoust	Univ. du Québec, Montréal	André Salem	Univ. Sorbonne Nouvelle Paris 3
Anne Dister	Univ. de Liège	Pascale Sébillot	IRISA / INSA Univ. de Rennes
Jules Duchastel	Univ. du Québec, Montréal	Max Silberstein	Univ. de Franche-Comté
Cédric Fairon	Univ. Catholique de Louvain	Ariana Tuzzi	Univ. di Padova
Serge Fleury	Univ. Sorbonne Nouvelle Paris 3	Jean-Marie Viprey	Univ. de Franche-Comté

**Organisation Committee**

Sergio Bolasco	Sapienza Univ. di Roma ( <i>President</i> )	Luca Giuliano	Univ. Roma 2 'Tor Vergata'
Francesco Baiocchi	Istat	Stella Iezzi	Sapienza Univ. di Roma
Simona Balbi	Univ. di Napoli 'Federico II'	Gevisa La Rocca	Univ. di Palermo
Isabella Chiari	Sapienza Univ. di Roma	Alessandra Leotta	Sapienza Univ. di Roma
Francesca Della Ratta	Istat	Bruno Mazzara	Sapienza Univ. di Roma
Francesca Dolcetti	Sapienza Univ. di Roma	Isabella Mingo	Sapienza Univ. di Roma
Augusto Frascatani	Sapienza Univ. di Roma	Arjuna Tuzzi	Univ. di Padova

**Referees**

Alessandra Areni, Enrica Aureli, Harald Baayen, Francesco Baiocchi, Simona Balbi, Barbara Baldazzi, Roberto Basili, Francois Bavaud, Valérie Beaudouin, Monica Becue, Sergio Bolasco, Giuseppe Bove, Martine Cadot, Paola Cavalieri, Isabella Chiari, Mirella Conenna, Michele Cortelazzo, François Daoust, Jean-Claude Deroubaix, Anne Dister, Francesca Dolcetti, Jules Duchastel, Ramón Alvarez Esteban, Cedric Fairon, Serge Fleury, Francesca Gamarotto, Kim Gerdes, Luca Giuliano, Benoit Habert, Serge Heiden, Stella Iezzi, Michel Jacobson, Margareta Kastberg, Mauro La Torre, Pierre Lafon, Ludovic Lebart, Jean-Marc Leblanc, Alain Lelu, Dominique Longrée, Pascal Marchand, William Martinez, Damon Mayaffre, Sylvie Mellet, Michelangelo Misuraca, Denis Monière, Hubert Naets, Rosamaria Paniccia, Bénédicte Pincemin, Gérald Purnelle, Max Reinert, Liliane Rodriguez, André Salem, Max Silberstein, Monique Slodzian, Maurice Tournier, Arjuna Tuzzi, Flavia Ursini, Jean-Marie Viprey, François Yvon

**Organised by:** Dip. di Studi Geoeconomici Linguistici Statistici Storici per l'Analisi Regionale, Sapienza Univ. di Roma

**With the support of:** SIS - Società Italiana di Statistica

**Sponsored by:**

**SAPIENZA**  
UNIVERSITÀ DI ROMA

Dip. di Studi Sociali Economici Attuariali Demografici  
Dip. di Sociologia e Comunicazione  
Dip. di Matematica e Statistica, Università di Napoli 'Federico II'  
Dip. di Sociologia, Università di Padova  
CISU (Casa Editrice)

**and by:**

Istat - Istituto Nazionale di Statistica

Enel - Ente Nazionale di Energia Elettrica

Percorsi srl

SAS Institute



**Graphic and Design:** Augusto Frascatani, Ida Potenza

**Artwork:** Caterina Bolasco

**Subjects:**

Statistical Methods, Exploratory Textual Data Analysis, Text Mining, Information Extraction, Text Categorization and Classification, Natural Language Processing, Computational Linguistics, Semantic Analysis, Content Analysis, Discourse Analysis

ISBN 978-88-7916-450-9

Copyright 2010

*LED* Edizioni Universitarie di Lettere Economia Diritto

Via Cervignano 4 - 20137 Milano

[www.lededizioni.com](http://www.lededizioni.com) - [www.ledonline.it](http://www.ledonline.it) - E-mail: [led@lededizioni.com](mailto:led@lededizioni.com)

All rights reserved for all countries.

No part of this work may be reproduced, stored electronically, or published in any form or by any means, digital or analog.

*Pageination and Layout:* Linda Cazzaniga

*Printed by:* Digital Print Service



# Contents

## Volume 1

13 ■ Preface

### PART I

#### METHODS AND EXPLORATORY TEXTUAL DATA ANALYSIS

##### **Automatic Classification and Indexing**

- 17 Inge Alberts, Dominic Forest, Suzanne Bertrand-Gastaldy  
*Catégorisation automatique du courriel en fonction de catégories pragmatiques*
- 27 Simona Balbi, Raffaele Miele, Germana Scepi  
*Clustering of documents from a two-way viewpoint*
- 37 Jean-François Chartier, Jean-Guy Meunier, Choukri Djellali  
*Analyse des variations entre partitions générées par différentes techniques de classification automatique de textes*
- 49 Jean Danis, Jean Guy Meunier, Jean-François Chartier, Motassem Alrahabi, Jean-Pierre Desclés  
*Classification automatique et stratégie d'annotation appliquées à un concept philosophique: la dimension psychologique du concept de LANGAGE dans l'œuvre de Bergson*
- 61 Luca Giuliano, Gevisa La Rocca  
*Validity and reliability of the automatic classification of texts according to the negative-positive criterion*
- 73 Rosita Guido, Michelangelo Misuraca, Francesca Vocaturo  
*An automatic SVM-based strategy for Digital Protocol*
- 83 Férihane Kboubi, Anja Habacha Chabi, Mohamed BenAhmed  
*L'exploitation des relations d'association de termes pour l'enrichissement de l'indexation de documents textuels*
- 93 Mauro La Torre, Susanna Pallini  
*Classificazione automatica di narrazioni autobiografiche*
- 105 Laurent Kevers, Amin Mantrach, Cédric Fairon, Hugues Bersini, Marco Saerens  
*Classification supervisée hybride par motifs lexicaux étendus et classificateurs SVM*
- 119 Mario A. Maggioni, Francesca Gambarotto, T. Erika Uberti  
*Mapping the evolution of industrial \*clusters\*: a meta-analysis*
- 131 Pasquale Pavone  
*Sintagmazione del testo: una scelta per disambiguare la terminologia e ridurre le variabili di un'analisi del contenuto di un corpus*
- 141 Dominique Peyrat-Guillard, Daniel Dufresne  
*L'analyse exploratoire de référentiels de compétences avec Alceste : une aide à la lecture de l'analyste*

### Text Mining

- 151 Carlo Amati, Fabio De Angelis, Francesca Romani  
*Textual data classification for a sectoral categorisation of public investments*
- 163 Vanessa Andreani, Thomas Lebarbé  
*Named entity normalization for termino-ontological resource design: mixing approaches for optimality*
- 173 Ismaïl Biskri, Hassane Hillali, Louis Rompré  
*Extraction de relations d'associations maximales dans les textes*
- 183 Martine Cadot, Michel Zitt, Gabriel Meurin, Alain Lelu  
*Robustesse des partitions de textes : une exploration autour de l'apport des motifs de mots*
- 195 Francesca della Ratta Rinaldi, Barbara Loré  
*Il lavoro e i suoi contenuti. Un'applicazione di Text Mining per categorizzare le attività dettagliate di lavoro nell'indagine campionaria sulle professioni Istat*
- 203 Hemant Misra, François Yvon  
*Modèles thématiques pour la segmentation de documents*
- 215 Juan-Manuel Torres-Moreno, Alejandro Molina, Gerardo Sierra  
*La energía textual como medida de distancia en agrupamiento de definiciones*

### Stylometry and Textometry

- 227 Delphine Amstutz, Philippe Gambette  
*Utilisation de la visualisation en nuage arboré pour l'analyse littéraire*
- 239 Anne Bandry-Scubbi  
*Du vocabulaire spécifique à l'analyse stylistique : l'exemple de Roderick Random*
- 253 Julien Bonneau, Anne Dister  
*Logométrie et modélisation des interactions discursives. L'exemple des entretiens semi-directifs*
- 265 Aurore Boulard, Jean-Marie Gauthier  
*Le complément sujet : Etude de l'utilisation des pronoms moi et je dans le discours d'enfants grâce à l'analyse statistique discursive*
- 275 Etienne Brunet  
*L'allitération. Hasard et observation*
- 289 Robert Byrnes  
*A statistical analysis of the «Eumaeus» phrasemes in James Joyce's Ulysses.*
- 297 Cyril Labbé, Dominique Labbé  
*Ce que disent leurs phrases*
- 309 Marc Le Pouliquen, Marc Csernel  
*Stemma codicum et relation d'intermédiarité, utilisation de la méthode de Don Quentin*
- 321 Xuan Luong, Etienne Brunet, Dominique Longrée, Damon Mayaffre, Sylvie Mellet, Céline Poudat  
*La cooccurrence, une relation asymétrique?*
- 333 Véronique Magri-Mourgues  
*Distance intertextuelle et connexion lexicale : outils de catégorisation générique ou stylistique ? Approche expérimentale d'un corpus inédit : le corpus aragonien*
- 341 Bénédicte Pincemin, Serge Heiden, Marie-Hélène Lay, Jean-Marc Leblanc, Jean-Marie Viprey  
*Fonctionnalités textométriques : Proposition de typologie selon un point de vue utilisateur*
- 355 Takafumi Suzuki, Shuntaro Kawamura, Akiko Aizawa  
*Exploratory analysis of stylistic characteristics in Japanese Q&A communities*

- 363 Takafumi Suzuki, Shuntaro Kawamura, Fuyuki Yoshikane, Kyo Kageura, Akiko Aizawa  
*Co-occurrence-based indicators for investigating authors' styles*
- 375 Xavier-Laurent Salvador, Fabrice Issac  
*Modèles théoriques inductifs et propositions d'applications aux données textuelles de l'ancien français*

### Lexical and Semantic Analysis

- 385 Yves Bestgen, Guy Lories, Jennifer Thewissen  
*Using latent semantic analysis to measure coherence in essays by foreign language learners?*
- 397 Jean-Pierre Colson  
*Automatic extraction of collocations: a new Web-based method*
- 409 Marie Pierre Escoubas-Benveniste  
*Le Monde et le « Grenelle de l'Environnement » : pistes pour l'analyse sémantique assistée par ordinateur d'un corpus de presse*
- 423 Margareta Kastberg Sjöblom  
*Costellazioni tematiche in un corpus letterario italiano*
- 433 Dominique Longrée, Caroline Philippart de Foy, Gérald Purnelle  
*Structures phrastiques et analyse automatique des données morphosyntaxiques : le projet LatSynt*
- 443 Fionn Murtagh, Adam Ganz  
*Semantics from Narrative*
- 455 Laure Pairet  
*Vocabulaire de la gestion: usage et sémantique*
- 467 Coralie Reutenauer, Evelyne Jacquy, Michelle Lecolle, Mathieu Valette  
*Sémème au microscope : genèse et variation sémiques d'une unité lexicale*
- 479 Irene Russo  
*Indicatori sintagmatici di bifunzionalità per gli aggettivi relazionali*
- 489 Jean-Marie Viprey, Philippe Schepens  
*Dérivation lexicale et dérive du discours : « mutualisation, mutualiser »*
- 499 Amal Zouaq, Michel Gagnon, Benoit Ozell  
*Grammaire de dépendances et ontologies de haut niveau : vers un processus modulaire pour l'analyse sémantique*

PART II  
DOMAIN SPECIFIC ANALYSIS

**Newspaper Analysis**

- 513 Alessandra Areni, Gilda Sensales, Alessandra Dal Secco  
*Le rappresentazioni del movimento del Sessantotto nella stampa italiana di quel periodo. Indagine lessicografica sui titoli di quotidiani di diverso orientamento ideologico-culturale*
- 525 Manuela Bussola, Federica Pellizzaro, Silvia Montecolle, Nicola Vallo, Ludovica Ioppolo, Fabio Marcodoppido, Federica Mancini, Paola Muccitelli, Manuela Nieddu, Francesca Proia  
*Le parole della disoccupazione nelle storie inviate a "La Repubblica". Un'integrazione tra gli approcci lessicometrico ed ermeneutico*
- 537 Joceline Chabot, Sylvia Kasparian, Philippe Desjardins  
*Massacres, atrocités et génocide. Analyse comparée d'un corpus de presse canadienne-française sur les atrocités allemandes et le génocide arménien (1914-1919)*
- 551 Alessandra Dal Secco, Gilda Sensales, Alessandra Areni  
*Representations of March 8 and feminine identities. A pilot study on communication in the Italian press (2000-2009)*
- 561 Alexandre Delanoë  
*Statistique textuelle et séries chronologiques sur un corpus de presse écrite. Le cas de la mise en application du principe de précaution*
- 573 Stefano Ondelli, Matteo Viale  
*Evidenze quantitative sull'italiano tradotto in un corpus giornalistico*
- 585 Cornelia Zuell  
*Using computer-assisted text analysis to identify media reported events*

**New Media and Social Networks**

- 597 Adil El Ghali, Yann Vigile Hoareau  
*Analysing the blogosphere using a random walk through its semantic spaces*
- 607 Cristina Cenci, Enrico Pozzi, Matteo Borsacchi  
*Autopsia semantica di un corpo mistico: la morte di Michael Jackson nella stampa italiana e su YouTube*
- 619 Louise-Amélie Cougnon, Thomas François  
*Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS*
- 631 Anna Gigante, Elisabetta Pelliccia  
*L'immagine delle parole in rete. Applicazione di Network Text Analysis sui gruppi di Facebook dedicati alla politica*
- 643 Daniel Hromada  
*Quantitative intercultural comparison by means of parallel pageranking of diverse wikipedias*
- 653 Jacques Savoy, Olena Zubareyva  
*Classification automatique d'opinions dans la blogosphère*

## Volume 2

### Legal, Scientific and Political Discourse

- 665 Lorenzo Bernardi, Arjuna Tuzzi  
*L'autografo del Presidente della Repubblica: un archetipo del discorso di fine anno mediante ADT*
- 677 Stefano Castelli, Alessandro Pepe, Loredana Addimando  
*Qualitatively mapping a research front through word-correspondence textual analysis: a case study*
- 685 Michele Cortelazzo, Francesca Gambarotto  
*I discorsi dei Presidenti di Confindustria*
- 697 Serge de Sousa  
*Parentages et proximités segmentales dans le discours révolutionnaire en Amérique latine de Bolivar à la Révolution bolivarienne (1811-2009)*
- 709 Silvia Fernandez, Pierre Jourlin, Eric SanJuan  
*Unsupervised mining of knowledge gaps in scientific literature*
- 719 Domenica Fioredistella Iezzi  
*Topic connections and clustering in text mining: an analysis of the JADT network*
- 731 Graeme Hirst, Yaroslav Riabinin, Jory Graham  
*Party status as a confound in the automatic classification of political speech by ideology*
- 743 Ersilia Incelli  
*Investigating keyword extraction for identifying units of stance in legislative texts*
- 755 Bill Louw, Carmela Chateau  
*Semantic prosody for the 21st Century: Are prosodies smoothed in academic contexts? A contextual prosodic theoretical perspective*
- 765 Erin MacMurray  
*Trois débats et une élection : Débats à l'occasion de l'élection présidentielle américaine de 2008 [Obama-McCain]*
- 779 Anne-Lyse Minard, Anne-Laure Ligozat, Brigitte Grau  
*Extraction de résultats expérimentaux d'articles scientifiques pour le peuplement d'une base de données*
- 791 Isabella Mingo, Cristina Panattoni  
*Il lessico e i temi della statistica ufficiale in Italia. Un'analisi lessicometrica del Programma Statistico Nazionale degli ultimi dieci anni*
- 805 Denis Monière, Dominique Labbé  
*Segmentation des corpus chronologiques : 143 ans de discours gouvernemental au Québec*
- 817 Maxime Sainte-Marie, Jean-Guy Meunier, Nicolas Payette, Jean-François Chartier  
*Reading Darwin between the lines: a computer-assisted analysis of the concept of evolution in the Origins of species*
- 827 Jacques Savoy  
*Discours électoral et discours présidentiel : Une étude lexicale comparative de B. Obama*
- 839 Ines Testoni , Arjuna Tuzzi, Marisa Cemin, Elisa Dakin  
*The bioethical debate between Laicism and Catholicism on the self-determination of death and dying. Gathering of logical substratum over and above opposites*



### Society and Psycho-/Socio-Linguistics

- 853 Priscille Baldit-Schneller, Catherine Dominguès  
*La carte de randonnée vue par les randonneurs*
- 865 Valérie Beaudouin, Julia Velkovska  
*Dialogues vocaux entre clients et automates ou comment l'homme et la machine s'entendent dans la réalisation d'un service*
- 877 Marcello Bidoli, Gevisa La Rocca, Chiara Mapelli  
*Erogare formazione in F.A.D. e in Blended. Individuazione delle problematiche dei tutor attraverso l'analisi dei segmenti*
- 885 Mathieu Brugidou, Michèle Moine  
*Normes émergentes et stigmatisation. Une analyse comparative à partir des deux questions ouvertes sur les raisons de ne pas trier les déchets et de ne pas faire d'économie d'énergie*
- 897 Pascual Cantos Gómez  
*Analyzing the oral speech of an Alzheimer affected person: A case study*
- 907 Grazia De Maio, Fiorenza Deriu  
*L'emergenza abitativa a Roma: dalla vulnerabilità all'esclusione sociale. Percorsi esistenziali*
- 917 Francesca della Ratta Rinaldi  
*Se pensa al suo futuro, di cosa ha più paura?*
- 929 Francesca della Ratta Rinaldi, Marianna De Luca  
*Le parole dei contratti. Quarant'anni di contrattazione in Enel: un'analisi sulle "premesse" e i "protocolli"*
- 939 Annette Gerstenberg  
*Analyse sociolinguistique d'un corpus oral par regroupement hiérarchique*
- 951 Eleonora Mussino, Laura Bernardi  
*Parlando di figli: analisi testuale delle aspettative di fecondità....*
- 963 Amélie Ngatsi-Imafouo  
*Analyse exploratoire d'un corpus d'apprentissage collaboratif: détection de reformulations multimodales lors des échanges enseignant-étudiant*
- 975 Julie Séguéla, Gilber Saporta, Stéphane Le Viet  
*e-Recrutement : recherche de mots-clés pertinents dans le titre des annonces d'emploi*
- 983 Gian Piero Turchi, Elena Celleghin, Martina Sarasin, Eleonora Pinto  
*Rappresentazione della realtà "sport" e della realtà "doping" attraverso un'analisi comparativa dei processi discorsivi praticati dalla categoria sportiva e dal "senso comune"*
- 997 Gian Piero Turchi, Luisa Orrù, Maria Sperotto, Sara Francato  
*La valutazione dell'efficacia di un intervento di mediazione attraverso la raccolta delle produzioni discorsive*

PART III  
APPLICATIONS AND TOOLS

**Software Applications and Tools**

- 1013 Fernande Dupuis, Robert Kapitan, François Daoust  
*Expérience d'entraînement de TreeTagger et d'intégration à l'interface Web de SATO*
- 1021 Serge Heiden, Jean-Philippe Magué, Bénédicte Pincemin  
*TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement*
- 1033 Nouha Chaâben Kammoun, Lamia Hadrich Belguith, Abdelmajid Ben Hamadou  
*The MORPH2 new version: A robust morphological analyzer for Arabic texts*
- 1045 Marie-Hélène Lay, Bénédicte Pincemin  
*Pour une exploration humaniste des textes : AnaLog*
- 1057 Jean-Marc Leblanc  
*Nouvelles fonctionnalités pour la visualisation des données textuelles et de résultats : Pour une approche plus ergonomique des dispositifs lexicométriques*
- 1069 Stefania Macchia, Manuela Murgia, Paola Vicari  
*Integration between automatic coding and statistical analysis of textual data systems*
- 1079 William Martinez, François Daoust, Jules Duchastel  
*Un service Web pour l'analyse de la cooccurrence*
- 1091 Mahsa Mohaghegh, Abdolhossein Sarrafzadeh  
*Performance evaluation of various training data in English-Persian Statistical Machine Translation*
- 1101 Mohammad Taher Pilevar, Heshaam Faili  
*PersianSMT: A first attempt to English-Persian Statistical Machine Translation*
- 1113 Juan-Manuel Torres-Moreno, Javier Ramirez Rodriguez  
*REG : Un algorithme glouton appliqué au résumé automatique de texte. Une approche exploratoire*
- 1123 Aris Xanthos, François Bavaud  
*Unsupervised learning of word separators with MDL*
- 1135 Xing Zhang, Alex Chengyu Fang  
*An ATE system based on probabilistic relations between terms and syntactic functions*

**Text Corpora Encoding**

- 1145 François Daoust, Yves Marcoux, Jean-Marie Viprey  
*L'annotation structurelle*
- 1157 Anne Dister, Gérald Purnelle, Richard Beaufort  
*Grands ou petits attentats ? Mesure de l'impact des réformes orthographiques sur la physionomie des textes*
- 1165 Annibale Elia, Fabiola Bocchino, Alberto Maria Langella, Mario Monteleone, Daniela Vellutino  
*Grammatiche locali per il riconoscimento automatico e la classificazione delle FAQ sull'Informazione Comunitaria Europea*
- 1175 Iris Eskhol, Isabelle Tellier, Samer Taalab, Sylvie Billot  
*Etiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques*

- 1187 Annamaria Landolfi, Carmela Sammarco, Miriam Voghera  
*Verbless clauses in Italian, Spanish and English: a Treebank annotation*
- 1195 Dominique Longrée, Sylvie Mellet, Céline Poudat  
*Les taggers, auxiliaires heuristiques en ADT ?*
- 1207 Maria Parascandolo, Francesco Cutugno.  
*A protocol to unify annotative standards on Italian treebanks*

### Monolingual and Multilingual Lexical Resources

- 1217 Steffen Eger, Ineta Sejane  
*Computing semantic similarity from bilingual dictionaries*
- 1227 Annibale Elia, Federica Marano, Mario Monteleone, Simona Sabatino, Daniela Vellutino  
*Strutture lessicali delle informazioni comunitarie all'interno di domini specialistici*
- 1237 Fabrice Issac  
*Outils et méthode de constitution de dictionnaire de formes figées*
- 1249 Mathieu Lafourcade, Alain Joubert  
*Construction de l'arbre des usages nommés d'un terme dans un réseau lexical évolutif*
- 1259 Gaël Lejeune, Nadine Lucas, Antoine Doucet  
*Tentative d'approche multilingue en extraction d'information*
- 1269 Ryo Nagata, Junichi Kakegawa, Takafumi Kutsuwa  
*Detecting missing sentence boundaries in learner English*
- 1277 Sophie Piron, Nadine Vincent  
*Un demi-siècle d'évolution des classements verbaux dans le Petit Larousse illustré*
- 1287 Gudrun Rawoens  
*Multilingual corpora in cross-linguistic research. Focus on the compilation of a Dutch-Swedish parallel corpus*
- 1295 Liliane Rodriguez  
*Le rôle du marquage lexicométrique des anglicismes dans un corpus franco-canadien*
- 1305 Fatiha Sadat  
*Using co-occurrence tendencies to improve Cross-Language Information Retrieval*
- 1317 Stefania Spina  
*AIWL: una lista di frequenza dell'italiano accademico*
- 1327 ■ Authors index

## Preface

The JADT (*Journées d'Analyse Statistique des Données Textuelles*) conference has become a reference point for all scholars interested in the methods and application of the statistical analysis of textual data. This year, it was held for the second time at the “La Sapienza” University of Rome (June 9-11, 2010). The meeting drew together a large group of statisticians, linguists, psychologists, sociologists and communication and IT experts, and was the tenth in the series – previous ones having been held in Barcelona (1990), Montpellier (1993), Rome (1995), Nice (1995), Lausanne (2000), Saint Malo (2002), Louvain-la-Neuve (2004), Besançon (2006) and Lyon (2008). Every two years, the JADT conference presents the state of the art for what concerns theories, problems, methods, algorithms and applications that share a ‘metric’ approach to the study of lexical, textual, pragmatic or discursive features of information expressed in natural language. The cover of the present book collects keywords extracted from the titles of the 800 papers presented so far at the JADT meetings.

The Proceedings of the 2010 conference, published in two volumes and, for the first time, on CD-ROM, collect 120 contributions by 250 scholars from 15 countries spread all over the world. Since the many applicative possibilities and problems of the statistical analysis of textual data correspond to a strong interdisciplinary character, these contributions are not easily classified. Nevertheless, an attempt has been made to help the reader by identifying three broad areas of interest: methods and exploratory textual data analysis; domain specific analysis; applications and tools.

The first part - METHODS AND EXPLORATORY TEXTUAL DATA ANALYSIS - deals with the use of original and innovative exploratory methodologies and analyses. Much attention has been given to the methods of automatic classification and indexing, ranging from the classification of autobiographical narratives or positive-negative evaluations to the comparison of different classification techniques, from SVM based strategies to Text Mining solutions. This section also includes papers on stylometry and textometry, two areas that share several measuring tools, interpretative criteria and areas of application. A focus on specific linguistic aspects (for instance, the automatic extraction of collocations, the measurement of textual coherence and the relationship between lexis, usage and semantics) characterizes the chapter on lexical and semantic analysis.

The second part - DOMAIN SPECIFIC ANALYSIS – covers the whole range of subjects that are explored by disciplines that make use of statistical analyses of textual data: for instance, the language of the press, the presentation of events by the media, and new communication technologies and applications in web 2.0 – such as social networks, blogs, forum and Wikipedia. The second volume begins with a chapter devoted to the analysis of legal, administrative, scientific and political discourse. Among other things, it includes studies of presidential discourses, analyses of semantic prosody in academic discourse, and an investigation of the lexis of the JADT network itself. This section ends with a series of contributions of a sociological and psycho-/socio-linguistic nature. These examine, for example, man-machine interactions, the spoken language of Alzheimer patients, and the lexis of trade union negotiations and job advertisements.

The third part presents research contributions on APPLICATIONS AND TOOLS. The first chapter of this section discusses software and tools for text mining and text analysis by means of web services, open-source platforms (TreeTagger, SATO, TXM) or automatic translation

tools for statistical applications. The second chapter is a collection of papers on the encoding and annotation of texts. It deals, in particular, with the difficulties involved in tagging spoken corpora, syntactic annotation, and annotation standards. Lastly, the third chapter collects works on the creation and use of multi- and mono-lingual lexical resources with a view to measuring similarities between bi-lingual dictionaries, creating dictionaries on idioms and building and using parallel corpora and frequency lists.

## Acknowledgements

We express our gratitude to the 58 reviewers who offered their invaluable assistance in the selection and anonymous refereeing of the papers in this volume: Alessandra Areni, Enrica Aureli, Harald Baayen, Simona Balbi, Barbara Baldazzi, Roberto Basili, Francois Bavaud, Valérie Beaudouin, Monica Becue, Sergio Bolasco, Giuseppe Bove, Martine Cadot, Paola Cavalieri, Isabella Chiari, Mirella Conenna, Michele Cortelazzo, François Daoust, Jean-Claude Deroubaix, Anne Dister, Francesca Dolcetti, Jules Duchastel, Ramón Álvarez Esteban, Cedrick Fairon, Serge Fleury, Francesca Gambarotto, Kim Gerdes, Luca Giuliano, Benoit Habert, Serge Heiden, Stella Iezzi, Michel Jacobson, Margareta Kastberg, Mauro La Torre, Pierre Lafon, Ludovic Lebart, Jean-Marc Leblanc, Alain Lelu, Dominique Longrée, Pascal Marchand, William Martinez, Damon Mayaffre, Sylvie Mellet, Michelangelo Misuraca, Denis Monière, Hubert Naets, Rosamaria Paniccia, Bénédicte Pincemin, Gérald Purnelle, Max Reinert, Liliane Rodriguez, André Salem, Max Silberztein, Monique Slodzian, Maurice Tournier, Arjuna Tuzzi, Flavia Ursini, Jean-Marie Viprey, François Yvon.

JADT2010 was held under the patronage of the SIS, Società Italiana di Statistica (Italian Statistics Society), and was funded by the “Sapienza” University of Rome and in particular, by the Department of Social, Economic, Demographic and Actuarial Studies and the Department of Sociology and Communication Science; by the Department of Mathematics and Statistics at the University of Naples ‘Federico II’; and by the Department of Sociology at the University of Padua. We are also very grateful to the following sponsors: Enel - Ente Nazionale di Energia Elettrica, Istat - Istituto Nazionale di Statistica, Percorsi S.r.l., and the SAS Institute. Not only have the above institutions contributed to the publication of the Proceedings, their in-house researchers have also made interesting and inspiring contributions to the conference as a whole.

As regards the organization of the conference, we would like to thank all the members of the local organizing committee and the staff of the Department of Geo-economic, Linguistic, Statistical and Historical Studies for Regional Analysis, Faculty of Economics. The following people deserve special mention: Rinaldo Coluccio, Paola D’Alonzo, Francesca Gargiulo, Alessandra Leotta, Patrizia Passacantilli, Raffaele Principe and Antonio Santini.

Special thanks go to Caterina Bolasco and Ida Potenza for providing the conference with a visual identity through their creative work; and to Francesco Baiocchi, Arianna Gattoni and Jonathan Anderlucci for their assistance with the web site.

For their precious help and painstaking revision and editing of the texts, we are grateful to Manuela Lo Prejato and Manuela Senza Peluso. We also wish to thank Giorgia Domanico and Marta Vincenzi for their editing assistance and secretarial work.

The editors