

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

PROPRIETÀ LETTERARIA RISERVATA

Copyright 2018 - UniversItalia - Roma

ISBN 978-88-3293-137-2

A norma della legge sul diritto d'autore e del codice civile è vietata la riproduzione di questo libro o di parte di esso con qualsiasi mezzo, elettronico, meccanico, per mezzo di fotocopie, microfilm, registratori o altro. Le fotocopie per uso personale del lettore possono tuttavia essere effettuate, ma solo nei limiti del 15% del volume e dietro pagamento alla SIAE del compenso previsto dall'art. 68, commi 4 e 5 della legge 22 aprile 1941 n. 633. Ogni riproduzione per finalità diverse da quelle per uso personale deve essere autorizzata specificatamente dagli autori o dall'editore.

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédric Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F

Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Toulouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I

Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingò: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Introduction XVII

Acknowledgements XIX

Invited Speakers

GERMAN KRUSZEWSKI

Memorize or generalize? Searching for a compositional RNN in a haystack

Adam Liška XXIII

BING LIU

Scaling-up Sentiment Analysis through Continuous Learning XXIV

PASCAL MARCHAND

La textométrie comme outil d'expertise :

application à la négociation de crise. XXV

GEORGE K. MIKROS

Author Identification Combining Various Author Profiles. Towards a Blended

Authorship Attribution Methodology XXVI

ROBERTO NAVIGLI

From text to concepts and back: going multilingual

with BabelNet in a step or two XXVII

Contributors

MOTASEM ALRAHABI¹, CHIARA MAINARDI¹

Identification automatique de l'ironie et des formes apparentées dans un

corpus de controverses théâtrales 1

MOHAMMAD ALSADHAN, SASCHA DIWERSY,

AGATA JACKIEWICZ, GIANCARLO LUXARDO

Migrants et réfugiés : dynamique de la nomination de l'étranger 10

R. ALVAREZ-ESTEBAN, M. BÉCUE-BERTAUT, B. KOSTOV, F. HUSSON, J-A

SÁNCHEZ-ESPIGARES

Xplortext, a R package. Multidimensional statistics for textual data science. 19

ELENA, AMBROSETTI, ELEONORA MUSSINO, VALENTINA TALUCCI

L'evoluzione delle norme: analisi testuale delle politiche sull'immigrazione in

Italia 26

MASSIMO ARIA, CORRADO CUCCURULLO

A bibliometric meta-review of performance measurement, appraisal, management research 35

LAURA ASCONE

Textual Analysis of Extremist Propaganda and Counter-Narrative: a quantitative investigation 44

LAURA ASCONE, LUCIE GIANOLA

Analyse de données textuelles appliquée à des problématiques de sécurité et d'enquête judiciaire 52

SIMONA BALBI, MICHELANGELO MISURACA, MARIA SPANO

A two-step strategy for improving categorisation of short texts 60

CHRISTINE BARATS, ANNE DISTER, PHILIPPE GAMBETTE, JEAN-MARC LEBLANC, MARIE PERES

Appeler à signer une pétition en ligne : caractéristiques linguistiques des appels 68

MANUEL BARBERA, CARLA MARELLO

Newsgroup e lessicografia: dai NUNC al VoDIM 76

IGNAZIA BARTHOLINI

Techniques for detecting the normalized violence in the perception of refugee / asylum seekers between lexical analysis and factorial analysis 83

PATRIZIA BERTINI MALGARINI, MARCO BIFFI, UGO VIGNUZZI

Dal corpus al dizionario: prime riflessioni lessicografiche sul Vocabolario storico della cucina italiana postunitaria (VoSCIP) 90

MARCO BIFFI

Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano postunitario 99

ANNICK FARINA, RICCARDO BILLERO

Comparaison de corpus de langue « naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues 108

FELICE BISOGNI, STEFANO PIRROTTA

Il rapporto tra famiglie di anziani non autosufficienti e servizi territoriali: un'analisi dei dati esplorativa con l'Analisi Emozionale del Testo (AET).... 117

ANTONELLA BITETTO, LUIGI BOLLANI

Esperienza di analisi testuale di documentazione clinica e di flussi informativi sanitari, di utilità nella ricerca epidemiologica e per indagare la qualità dell'assistenza 126

GUIDO BONINO, DAVIDE PULIZZOTTO, PAOLO TRIPODI

Exploring the history of American philosophy in a computer-assisted framework 134

MARC-ANDRE BOUCHARD, SYLVIA KASPARIAN

La classification hiérarchique descendante pour l'analyse des représentations sociales dans une pétition antibilinguisme au Nouveau-Brunswick, Canada 142

LIVIA CELARDO, RITA VALLEROTONDA, DANIELE DE SANTIS, CLAUDIO SCARICI, ANTONIO LEVA

Analysing occupational safety culture through mass media monitoring..... 150

BARBARA CORDELLA, FRANCESCA GRECO, PAOLO MEOLI, VITTORIO PALERMO, MASSIMO GRASSO

Is the educational culture in Italian Universities effective? A case study..... 157

MICHELE A. CORTELAZZO, GEORGE K. MIKROS, ARJUNA TUZZI

Profiling Elena Ferrante: a Look Beyond Novels 165

FABRIZIO DE FAUSTI, MASSIMO DE CUBELLIS, DIEGO ZARDETTO¹

Word Embeddings: a Powerful Tool for Innovative Statistics at Istat 174

Gibbons A. (1985). *Algorithmic Graph Theory*. Cambridge University Press. . 182

VIVIANA DE GIORGI, CHIARA GNESI

Analisi di dati d'impresa disponibili online: un esempio di data science tratto dalla realtà economica dei siti di e-commerce 183

ALESSANDRO CAPEZZUOLI, FRANCESCA DELLA RATTA, STEFANIA MACCHIA, MANUELA MURGIA, MONICA SCANNAPIECO, DIEGO ZARDETTO

The use of textual sources in Istat: an overview..... 192

FRANCESCA DELLA RATTA, GABRIELLA FAZZI, MARIA ELENA PONTECORVO, CARLO VACCARI, ANTONINO VIRGILLITO

Twitter e la statistica ufficiale: il dibattito sul mercato del lavoro 200

SAMI DIAF

Gauging An Author's Mood Using Hidden Markov Chains 209

MARC DOUGUET

Les hémistiches répétés 215

FRANCESCA DRAGOTTO, SONIA MELCHIORRE

«Mangiata dall'orco e tradita dalle donne». Vecchi e nuovi media raccontano la vicenda di Asia Argento, tra storytelling e Speech Hate 223

CRISTIANO FELACO, ANNA PAROLA

Il *cosa* e il *come* del processo narrativo. L'uso combinato della Text Analysis e Network Text Analysis al servizio della precarietà lavorativa 233

ANA NORA FELDMAN

Hablando de crisis: las comunicaciones del Fondo Monetario Internacional 242

VALERIA FIASCO

Brexit in the Italian and the British press: a bilingual corpus-driven analysis 250

VIVIANA FINI, GIUSEPPE LUCIO GAETA, SERGIO SALVATORE

Textual analysis to promote innovation within public policy evaluation 259

ALESSIA FORCINITI, SIMONA BALBI

- A proposal for Cross-Language Analysis:
violence against women and the Web 268

BEATRICE FRACCHIOLLA, OLINKA SOLENE DE ROGER

- La verbalisation des émotions 276

**LUISA FRANCHINA, FRANCESCA GRECO, ANDREA LUCARIELLO,
ANGELO SOCIAL, LAURA TEODONNO**

- Improving Collection Process for Social Media Intelligence: A Case Study . 285

**ANDREA FRONZETTI COLLADON, JOHANNE SAINT-CHARLES, PIERRE
MONGEAU**

- The impact of language homophily and similarity of social position on
employees' digital communication 293

MATTEO GERLI

- Looking Through the Lens of Social Sciences: The European Union in the EU-
Funded Research Projects Reporting 300

LUCIE GIANOLA, MATHIEU VALETTE

- Spécialisation générique et discursive d'une unité lexical L'exemple de
joggeuse dans la presse quotidienne régionale 312

PETER A. GLOOR, JOAO MARCOS DE OLIVEIRA, DETLEF SCHODER

- The Transparency Engine – A Better Way to Deal with Fake News 319

FRANCESCA GRECO, LEONARDO ALAIMO, LIVIA CELARDO

- Brexit and Twitter: The voice of people..... 327

FRANCESCA GRECO, GIULIO DE FELICE, OMAR GELO

- A text mining on clinical transcripts of good and poor outcome
psychotherapies 335

FRANCESCA GRECO, DARIO MASCHIETTI, ALESSANDRO POLLI

- DOMINIO: A Modular and Scalable Tool for the Open Source Intelligence 343

LEONIE GRÖN, ANN BERTELS, KRIS HEYLEN

- Is training worth the trouble? A PoS tagging experiment with Dutch clinical
records..... 351

FRANCE GUERIN-PACE, ELODIE BARIL

- Les outils de la statistique textuelle pour analyser
les corpus de données d'enquêtes de la statistique publique 359

SERGE HEIDEN

- Annotation-based Digital Text Corpora Analysis within the TXM Platform 367

DANIEL HENKEL

- Quantifying Translation : an analysis of the conditional perfect in English-
French comparable-parallel corpus..... 375

DANIEL DEVATMAN HROMADA

- Extraction of lexical repetitive expressions from complete works of William
Shakespeare..... 384

OLIVIER KRAIF, JULIE SORBA

Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale) 392

CYRIL LABBE, DOMINIQUE LABBE

Les phrases de Marcel Proust 400

LUDOVICA LANINI, MARÍA CARLOTA NICOLÁS MARTÍNEZ

Verso un dizionario *corpus-based* del lessico dei beni culturali: procedure di estrazione del lemmario 411

DANIELA LARICCHIUTA, FRANCESCA GRECO, FABRIZIO PIRAS, BARBARA CORDELLA, DEBORA CUTULI, ELEONORA PICERNI, FRANCESCA ASSOGNA, CARLO LAI, GIANFRANCO SPALLETTA, LAURA PETROSINI

"The grief that doesn't speak": Text Mining and Brain Structure 419

GEVISA LA ROCCA, CIRUS RINALDI

Icone gay: tra processi di normalizzazione e di resistenza. Ricostruire la semantica degli hashtag..... 428

LUDOVIC LEBART

Looking for *topics*: a brief review..... 436

GAËL LEJEUNE, LICHAO ZHU

Analyse Diachronique de Corpus : le cas du poker..... 444

JULIEN LONGHI, ANDRE SALEM

Approche textométrique des variations du sens..... 452

LAURENT VANNI¹, DAMON MAYAFFRE, DOMINIQUE LONGREE

ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables 459

LUCIE LOUBERE

Déconstruction et reconstruction de corpus... À la recherche de la pertinence et du contexte 467

HEBA METWALLY

L'apport du *corpus-maquette* à la mise en évidence des niveaux descriptifs de la chronologie du sens. Essai sur une Série Textuelle Chronologique du *Monde diplomatique* (1990-2008). 474

JUN MIAO, ANDRE SALEM

Séries textuelles homogènes..... 491

SILVIO MIGLIORI, ANDREA QUINTILIANI, DANIELA ALDERUCCIO, FIORENZO AMBROSINO, ANTONIO COLAVINCENZO, MARIALUISA MONGELLI, SAMUELE PIERATTINI, GIOVANNI PONTI SERGIO BOLASCO, FRANCESCO BAIOCCHI, GIOVANNI DE GASPERIS

TaLTaC in ENEAGRID Infrastructure..... 501

ISABELLA MINGO, MARIELLA NOCENZI

The dimensions of Gender in the International Review of Sociology. A lexicometric approach to the analysis of the publications in the last twenty years 509

ADIEL MITTMANN, ALCKMAR LUIZ DOS SANTOS

The Rhythm of Epic Verse in Portuguese From the 16th to the 21st Century 514

DENIS MONIERE, DOMINIQUE LABBE

Le vocabulaire des campagnes électorales 522

CYRIELLE MONTRICHARD

Faire émerger les traces d'une pratique imitative dans la presse de tranchées à l'aide des outils textométriques 532

ALBERT MORALES MORENO

Evolución diacrónica de la terminología y la fraseología jurídico-administrativa en los Estatutos de autonomía de Catalunya de 1932, 1979 y 2006 541

CEDRIC MOREAU

Comment penser la recherche d'un signe pour une plateforme multilingue et multimodale français écrit / langue des signes française ? 556

JEAN MOSCAROLA, BORIS MOSCAROLA

Conclusion ADT et visualisation, pour une nouvelle lecture des corpus Les débats de 2ème tour des Présidentielles (1974-2017) 563

MAURIZIO NALDI

A conversation analysis of interactions in personal finance forums 571

STEFANO NOBILE

Analisi testuale, rumore semantico e peculiarità morfosintattiche: problemi e strategie di pretrattamento di corpora speciali 578

DANIEL PELISSIER

L'individu dans le(s) groupe(s) : focus group et partitionnement du corpus 586

BENEDICTE PINCEMIN, CELINE GUILLOT-BARBANCE, ALEXEI**LAURENTIEV**

Using the First Axis of a Correspondence Analysis as an Analytical Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts 594

CELINE POUDAT

Explorer les désaccords dans les fils de discussion du Wikipédia francophone 602

MATTHIEU QUIGNARD, SERGE HEIDEN, FREDERIC LANDRAGIN,**MATTHIEU DECORDE**

Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes 610

PIERRE RATINAUD

Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ 616

LUISA REVELLI

- Il parametro della *frequenza* tra paradossi e antinomie:
il caso dell'*italiano scolastico* 626

PIERGIORGIO RICCI

- How Twitter emotional sentiments mirror on the Bitcoin
transaction network 635

CHANTAL RICHARD, SYLVIA KASPARIAN

- Analyse de contenu versus méthode Reinert : l'analyse comparée d'un corpus
bilingue de discours acadiens et loyalistes du N.-B., Canada 643

VALENTINA RIZZOLI, ARJUNA TUZZI

- Bridge over the ocean: Histories of social psychology in Europe and North
America. An analysis of chronological corpora 651

LOUIS ROMPRE, ISMAÏL BISKRI

- Les « itemsets fréquents » comme descripteurs de documents textuels 659

CORINNE ROSSARI, LJILJANA DOLAMIC, ANNALENA HÜTSCH, CLAUDIA RICCI, DENNIS WANDEL

- Discursive Functions of French Epistemic Adverbs: What can Correspondence
Analysis tell us about Genre and Diachronic Variation? 668

VANESSA RUSSO, MARA MARETTI, LARA FONTANELLA, ALICE TONTODIMAMMA

- Misleading information in online propaganda networks 676

ELIANA SANANDRES, CAMILO MADARIAGA, RAIMUNDO ABELLO

- Topic modeling of Twitter conversations 684

FRANCESCO SANTELLI, GIANCARLO RAGOZINI, MARCO MUSELLA

- What volunteers do? A textual analysis of voluntary activities in the Italian
context 692

S. SANTILLI, S. SBALCHIERO, L. NOTA, S. SORESI

- A longitudinal textual analysis of abstract presented at Italian Association for
Vocational guidance and Career Counseling'
Conferences from 2002 to 2017 700

JACQUES SAVOY

- A la poursuite d'Elena Ferrante 707

JACQUES SAVOY

- Regroupement d'auteurs dans la littérature du XIXe siècle 716

STEFANO SBALCHIERO, ARJUNA TUZZI

- What's Old and New? Discovering Topics in the American Journal of
Sociology 724

NILS SCHAETTI, JACQUES SAVOY

- Comparison of Neural Models for Gender Profiling 733

LIONEL SHEN

- Segments répétés appliqués à l'extraction de connaissances trilingues 740

SANDRO STANCAMPIANO	
Misurare, Monitorare e Governare le città con i Big Data	748
FADILA TALEB, MARYVONNE HOLZEM	
Exploration textométrique d'un corpus de motifs juridiques dans le droit international des transports	755
JAMES M. TEASDALE	
The Framing of the Migrant: Re-imagining a Fractured Methodology in the Context of the British Media.	763
MARJORIE TENDERO¹, CECILE BAZART	
Results from two complementary textual analysis software (Iramuteq and Tropes) to analyze social representation of contaminated brownfields	771
MATTEO TESTI, ANDREA MERCURI, FRANCESCO PUGLIESE	
Multilingual Sentiment Analysis.....	780
JUAN MARTÍNEZ TORVISCO	
A linguistic analysis of the image of immigrants' gender in Spanish newspapers.....	788
FRANCESCO URZÌ	
Lo strano caso delle frequenze zero nei testi legislativi euroistituzionali.....	796
SYLVIE VANDAELE	
Les traductions françaises de <i>The Origin of Species</i> : pistes lexicométriques .	805
PIERRE WAVRESKY, MATTHIEU DUBOYS DE LABARRE, JEAN-LOUP LECOEUR	
Circuits courts en agriculture : utilisation de la textométrie dans le traitement d'une enquête sur 2 marchés	814
MARIA ZIMINA, NICOLAS BALLIER	
On the phraseology of spoken French: initial salience, prominence and lexicogrammatical recurrence in a prosodic-syntactic treebank <i>Rhapsodie</i>	822

Abstracts

FILIPPO CHIARELLO, GUALTIERO FANTONI, ANDREA BONACCORSI, SILVIA FARERI	
What kind of contributions does research provides? Mapping issue based statements in research abstracts	833
FILIPPO CHIARELLO, GIACOMO OSSOLA, GUALTIERO FANTONI, ANDREA BONACCORSI, ANDREA CIMINO, FELICE DELL'ORLETTA	
Technical sentiment analysis: predicting the success of new products using social media.....	835

FIorenza DERIU, DOMENICA FIOREDISTELLA IEZZI
 Citizens and neighbourhood life: mapping population sentiment in Italian cities..... 837

FRANCESCA DI CARLO, ROSY INNARELLA, BRIZIO LEONARDO TOMMASI
 Vax network: profiling influential nodes with social network analysis on twitter..... 838

DAVIDE DONNA
 Alteryx 840

VALERIO FICCADENTI, ROY CERQUETI, MARCEL AUSLOOS
 Complexity of US President Speeches 841

PETER A. GLOOR
 Measuring the Dynamics of Social Networks with Condor 842

IOLANDA MAGGIO, DOMENICA FIOREDISTELLA IEZZI, MATTEO FATIGHENTI
 "BIG DATA" Words Trend Analysis using the multidimensional analysis of texts 844

MARIO MASTRANGELO
 Itinerari turistici, network analysis e text mining 845

MARIA FRANCESCA ROMANO, GUIDO REY, ANTONELLA BALDASSARINI PASQUALE PAVONE
 Text Mining per l'analisi qualitativa e quantitativa dei dati amministrativi utilizzati dalla Pubblica Amministrazione..... 847

ALESSANDRO CESARE ROSA
 Taglio cesareo e Vbac in Italia al tempo dei Big Data: una proposta di ulteriore contributo informativo..... 849

A text mining on clinical transcripts of good and poor outcome psychotherapies

Francesca Greco¹, Giulio de Felice², Omar Gelo³

¹Sapienza University of Rome & Prisma S.r.l. – francesca.greco@uniroma1.it

²Sapienza University of Rome & NCU University – giulio.defelice@uniroma1.it

³University of Salento & Sigmund Freud University – omar.gelo@unisalento.it

Abstract

The text mining of clinical transcripts is broadly used in psychotherapy research, but is limited to top-down approaches, with a-priori vocabularies that code the transcripts according to a theoretical predetermined framework. Nevertheless, the semantic level that a word or clinical intervention can assume depends on the relational field in which the discourse is produced. Thus, bottom-up approaches seem to be particularly meaningful in addressing such a relevant issue. With the aim of investigating possible similarities and differences between good outcome and poor outcome psychotherapies, we applied a multivariate analysis on the transcripts of eight single cases of brief experiential psychotherapy (four good outcome vs four poor outcome cases), in order to identify the general core themes, and their difference according to therapy outcome. The results showed a significant difference in the number of context units classified in two of the six core themes (clusters) between good and poor outcome cases (χ^2 , $df=5$, $p<0,01$). These findings show how the bottom-up technique of text analysis on clinical transcripts turned out to be an enlightening tool to let their latent dimensions emerge, setting the clinical process and outcome, and therefore, providing a very useful tool for clinical purposes.

Abstract

L'analisi delle trascrizioni cliniche è stata ampiamente utilizzata nella ricerca in psicoterapia, sebbene prevalentemente si basi sull'utilizzo di un dizionario che consente la codifica del testo in funzione di criteri predeterminati. Tuttavia, la polisemia che una parola, o un intervento clinico, può assumere dipende dal campo relazionale in cui il discorso è prodotto. Pertanto, gli approcci bottom-up sembrano essere particolarmente utili nell'affrontare tale questione. Allo scopo di indagare gli elementi caratterizzati le trascrizioni cliniche con esito positivo e negativo, è stata effettuata un'analisi multivariata di un corpus composto da otto trascrizioni di psicoterapia breve (quattro con esito positivo e quattro con esito negativo) al fine di identificare i temi centrali generali e la distribuzione delle unità di contesto nei diversi temi in

funzione dell'esito della terapia. I risultati hanno evidenziato una differenza significativa tra i casi con esito positivo e quelli con esito sfavorevole (χ^2 , $df = 5$, $p < 0,01$), mettendo in evidenza come l'analisi automatica del testo delle trascrizioni dei colloqui clinici possa essere uno strumento utile a far emergere le dimensioni latenti organizzatrici del processo e del risultato, configurandosi così come un utile strumento ai fini clinici.

Keywords: Emotional Text Mining, clinical transcripts, psychotherapy outcome.

1. Introduction

The text mining of clinical transcripts is very broadly used in psychotherapy research, but is limited to top-down approaches where *a-priori* vocabularies code them according to a theoretical predetermined framework. Nevertheless, the semantic level that a word, or clinical intervention, can assume, depends on the relational field in which the discourse is produced. Thus, bottom-up approaches seem to be particularly meaningful in addressing such relevant issue. Psychotherapy can be considered a dynamic communicative exchange between the client and the therapist (e.g., Gelo et Salvatore, 2016). Within such an exchange, the content (i.e., the semantic) of what is said plays a primary role. Thus, the textual analysis of therapy transcripts may represent a very useful tool for psychotherapy process researchers as well as for clinicians (Gelo et al., 2013; Salvatore et al. 2017). In the field of psychotherapy research, some methods of text mining have been developed and applied, such as the Therapeutic Cycle Model (Mergenthaler, 2008) and Referential Activity (Bucci et al., 1992). Following a *top-down* approach, these methods use predefined content categories to semantically classify units of text. Each of these categories corresponds to a thematic dictionary containing all the words indicative of the content represented by that category. Even though these top-down methods of text mining allow for a reliable and valid investigation of the therapeutic process, they present a major limitation, disregarding the contextual nature of the linguistic meaning (Carli et al., 2004; Salvatore et al., 2012). In fact, the meaning of a word is polysemic and depends on the way it combines with other words in the communicative interaction, i.e., it depends on its association with other words. Grounded on these considerations, there has recently been a development of text mining approaches which, by means of their bottom-up logic, allow for a context-sensitive textual analysis (e.g., Salvatore et al., 2012; 2017; Cordella et al., 2014; Greco, 2016). The aim of this study is to investigate possible similarities and differences between good outcome and poor outcome psychotherapy cases by applying the Emotional Text Mining (Cordella et al., 2014; Greco, 2016). Our assumption is that it is possible to

detect the associative links between the words in order to infer the symbolic matrix determining the coexistence of the terms in the text. To this aim, we perform a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The interpretation of the cluster analysis allows for the identification of the elements characterizing the core themes of the treatment, while the results of the correspondence analysis reflect the emotional symbolisation characterising the therapeutic exchange. The advantage of such an approach is to interpret the factorial space according to word polarization, thus identifying the emotional categories that generate the core themes, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space (Greco et al., 2017).

2. Data collection and analysis

2.1. Data collection

The sample of the present study was drawn from the York Depression Study I, a randomized clinical trial to assess the efficacy of brief experiential therapy for depression (Greenberg et Watson, 1998; Watson et al., 1998).¹ From the original sample, we initially selected the six best outcome cases and the six cases worst outcome cases based on the Reliable Change Index of the Beck Depression Inventory (BDI; Beck et al., 1988). We then excluded four cases due to missing session transcripts. Our final sample was thus comprised of a total of eight cases, with four good outcomes and four poor outcomes. The treatment length was between 15 and 20 sessions ($M = 17.62$; $SD = 1.38$), for a total of 141 sessions. Patients (one man and seven women; $M=37.1$ years old) met the criteria for major depressive disorder assessed by means of the Structured Clinical Interview for DSM-III-R (SCID; Spitzer et al., 1989). Therapists (seven women and one man; $M= 5.5$ years of therapeutic experience) had six months of training in experiential psychotherapy (Greenberg et al., 1993). The transcripts were collected in a large size corpus of 1090234 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the percentage of hapax ($TTR = 0.01$; $\text{hapax} = 35.3\%$). They highlighted the richness of the corpus indicating the possibility of proceeding with the analysis.

¹ We are grateful to Dr. Les Greenberg for having us provided with files of the transcripts for these cases.

2.2. Data analysis

First, data were cleaned and pre-processed with the software T-Lab and keywords selected. In particular, we used lemmas as keywords instead of type. We selected all the lemmas in the medium rank of frequency (upper frequency threshold = 933), and those of the low rank of frequency until the threshold of 17 occurrences, that is, equal to the average number of sessions made on average by the patients (Greco, 2016). Then, in order to identify the core themes common to all the psychotherapies, we performed a cluster analysis on the keywords per context units (CU) matrix, by means of a bisecting *k*-means algorithm (Savaresi et Boley, 2004), limited to ten partitions, excluding all the CU that did not have at least two keywords co-occurrences. The eta squared value was used to evaluate and choose the optimal solution. To finalize the text mining, we performed a correspondence analysis on the keywords per clusters matrix (Lebart et Salem, 1994) in order to explore the relationship between clusters, and to identify the emotional categories setting the psychotherapeutic process. The interpretation of the factorial space was performed according to the procedure proposed by Cordella and colleagues (2014) in which each keyword is considered only in the factor with the greatest absolute value. To finalise the analysis, we performed a chi squared test on the contingency table cluster per therapy outcome, calculating the standard residual in order to identify the differences between good outcome and poor outcome clinical transcripts in terms of core themes.

3. Main results and discussion

The results of the cluster analysis show that the 1351 keywords selected allow for the classification of 56.6% of context units. The high proportion of unclassified context units is due to the transcripts richness in terms of paraverbal interactions (i.e. *mhm*, *yeah*, etc). The eta squared value was calculated on partitions from 3 to 9, and it showed six clusters as the optimal solution ($\eta^2 = 0.034$). In table 1, we can appreciate the emotional map emerging from the clinical transcripts representing the clusters location in the factorial space produced by the interpretation of the five factors. The first factor reflects patient *positioning*, which can be passive or active; the second factor refers to the *relationship* that could be familiar or unfamiliar, i.e., a person facing something new and unpredictable; the third factor represents the *communication content* that can be emotional or concrete; the fourth factor reflects the *outcome* of the therapeutic work, that is, the patient's empowerment or making sense of the patient's experiences; and the fifth factor distinguishes the *issues* within the daily ones, concerning everyday life,

from the relational ones, concerning the loved ones.²

Table 1 – Factorial space representation (the percentage of explained inertia is reported between brackets under each factor).

Cluster	Label (CU%)	Factor 1 (26.7%) Positioning	Factor 2 (25.8%) Relationship	Factor 3 (21.5%) Content	Factor 4 (14.5%) Outcome	Factor 5 (11.5%) Issues
1	Family Structure (11.6%)	Passive 0.20	Familiar -0.56	Emotional -0.16		Daily -0.32
2	Transformative Process (12.1%)	Active -0.46	Unfamiliar 0.29	0.06	To empower -0.35	Daily -0.16
3	Concrete thinking (16.1%)	Passive 0.84	Unfamiliar 0.34	Concrete 0.42	To empower -0.19	0.05
4	Therapeutic Relationship (22.4%)	Active -0.25	Familiar -0.18	Concrete 0.41	To understand 0.28	Relational 0.16
5	Relational Issues (14.6%)	0.04	Familiar -0.14	Emotional -0.47	To empower -0.18	Relational 0.45
6	Feelings (23.1%)	0.06	Unfamiliar 0.58	Emotional -0.43	To understand 0.49	Daily -0.14

CU = context units classified in the cluster.

Table 2 – Psychotherapy core themes.

Cluster 1 Family Structure		Cluster 2 Transformative Process		Cluster 3 Concrete Thinking		Cluster 4 Therapeutic Relationship		Cluster 5 Relational Issues		Cluster 6 Feelings	
keyword	CU	keyword	CU	keyword	CU	keyword	CU	keyword	CU	keyword	CU
home	525	start	507	hear	455	week	699	mother	399	understand	416
kid	371	able to	504	money	326	sense	675	life	335	hurt	300
house	290	change	438	dollar	267	day	438	problem	333	important	298
father	241	different	396	accept	205	bad	432	hard	292	person	231
husband	213	situation	288	pay	196	angry	381	care	268	hard	213
child	205	point	237	listen	175	call	253	deal	252	support	185
parent	194	go on	216	believe	135	night	189	family	237	inside	170
stay	190	mind	213	matter	130	morning	169	relationship	233	strong	168
live	179	trying	183	sell	126	set	162	Father	195	pain	153

CU = context units classified in the cluster.

The six clusters are of different sizes (table 1) and reflect the core themes of the brief psychotherapy (table 2). The first cluster describes the *family structure* with its role and places; the second cluster reflects the *transformative*

² In the negative pole of the fifth factor (Daily Issues) we find the following words: *house, stay, TV, rule, street, teacher, move out, neighbour, pounds*, and in the positive pole we find words as *mother, life, problem, sister, relationship*.

process characterising a psychotherapy; the third cluster highlights the *concrete thinking* process, a way to think that could be defined as concrete thinking, which is often rational and frequently concerning economic issues; the fourth cluster represents the *therapeutic relationship* that is made of concrete limits, and the process of making sense of personal experiences; the fifth cluster reflects the *relational issues* of the patient's private life; and the sixth cluster refers to the process of detecting, recognizing, and understanding *feelings*, characterizing internal emotional experiences.

There is a significant difference in the number of content units classified in each cluster among the good and poor outcome therapies (χ^2 , $df = 5$, $p < 0.01$). In particular, the differences lay on the relevance of two of the six core themes: the *concrete thinking* and the *feelings*. While the good outcome brief psychotherapies are characterized by a high number of context units classified in the cluster *feelings* ($SE = 6.8$) and a low number of context units classified in the cluster *concrete thinking* ($SE = -5.8$); the poor outcomes psychotherapies are characterized by a high number of context units classified in the cluster *concrete thinking* ($SE = 6.8$) and a low number of context units classified in the cluster *feelings* ($SE = -7.0$). Namely, it would seem that patients tend to dwell upon their emotional experiences in the good outcome psychotherapy, while they tend to dwell upon facts in the poor outcome psychotherapy, probably not connecting them to their emotional experiences. Given that we classified the interactions among the patients and the therapists in this analysis, the therapy outcome could derive both from the patient's ability in dealing with feelings or the therapist's ability to support the patient in doing so.

The above-mentioned differences between good and poor outcome cases are coherent with findings obtained on the same sample by means of a principal component analysis made on the transcripts coded according to three dictionaries: abstract language, emotional positive language, and emotional negative language (de Felice et al., 2018). In this study, differences in the correlation matrices between good outcome and poor outcome cases were evident. The most obvious one concerned the dynamic in which the patient made use of abstract/concrete language, interpreted very positively in poor outcome cases and very negatively in good outcome cases. In the latter, it was probably and correctly considered as a patient's defense mechanism to address. This was confirmed by the use of positive and negative emotional language, inversely proportional to abstraction, only in poor outcome cases.

4. Conclusion

Talking about concrete events without any sort of emotional involvement in the clinical literature is a defence mechanism that goes under the name of

rationalisation, and it represents a way to protect the mind from painful feelings using an abstract, intellectual and often concrete attitude in dealing with them. While the good outcome psychotherapeutic relationships seem to be capable of addressing the emotional content laying under the surface of the psychotherapeutic field (i.e. use of the therapist's negative emotional language), the poor outcome dynamics seem to be completely wrapped up in a process of avoiding it. Both the PCA (de Felice et al 2018) and text analysis on clinical transcripts confirmed the difficulty in poor outcome psychotherapies to work on the patient's emotional aspects. This bottom-up technique of text analysis on clinical transcripts turned out to be an enlightening tool to let their latent dimensions emerge, arranging the clinical process and outcome, therefore, providing a very useful tool for clinical purposes.

References

- Beck A.T., Steer R.A. and Garbin M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8: 77-100.
- Bucci W., Kabasakalian-McKay R. and RA Research Group (1992). Scoring referential activity. Ulm, Germany: Ulmer Textbank.
- Carli R., Dolcetti F. and Dolcetti (2004). L'Analisi Emozionale del Testo (AET): un caso di verifica nella formazione professionale. In Purnelle G., Fairon C. and Dister A., editors, *Actes JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles*, pp. 250-261.
- 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 Nee E., Daube M., Valette M. and Fleury S., editors, *Actes JADT 2014 (12es Journées internationales d'Analyse Statistique des Données Textuelles, Paris, France, Juin 3-6, 2014)*, pp. 173-184.
- de Felice G., Orsucci F., Mergenthaler E., Gelo O., Paoloni G., Scozzari A., Serafini G., Andreassi S., Vegni N. and Giuliani A. (2018). What differentiates good and poor outcome psychotherapies? A statistical mechanics approach to psychotherapy research. *Nonlinear Dynamics, Psychology and Life Sciences*. Submitted.
- Gelo O.C.G. and Salvatore S. (2016). A dynamic systems approach to psychotherapy: A meta-theoretical framework for explaining psychotherapy change processes. *Journal of Counseling Psychology*, 63(4): 379-395.
- Gelo O.C.G., Salcuni S. and Colli A. (2013). Text analysis within quantitative and qualitative psychotherapy process research: introduction to special issue. *Res. Psychother. Psychopathol. Process Outcome* 15: 45-53.

- Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.
- Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *Rivista Italiana di Economia Demografia e Statistica*, 71(2): 125:36.
- Greenberg L., Rice L. and Elliott R. (1993). *Facilitating emotional change. The moment by moment process*. Guilford Press.
- Greenberg LS, Watson JC (1998). Experiential therapy of depression: differential effects of client-centered relationship conditions and process experiential interventions. *Psychotherapy-Research* 8: 210-224.
- Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod
- Mergenthaler E. (2008). Resonating minds: A school-independent theoretical conception and its empirical application to psychotherapeutic processes. *Psychotherapy Research*, 18(2): 109-126.
- Salvatore S., Gelo O., Gennaro A., Metrangolo R., Terrone G., Pace V., Venuleo C., Venezia A. and Ciavolino E. (2017). An automated method of content analysis for psychotherapy research: A further validation. *Psychotherapy Research*, 27(1): 38-50.
- Salvatore S., Gennaro A., Auletta A.F., Tonti M. and Nitti M. (2012). Automated method of content analysis: A device for psychotherapy process research. *Psychotherapy Research*, 22(3): 256-273.
- Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.
- Spitzer R., Williams J., Gibbons M. and First M. (1989). *Structured Clinical Interview for DSM-III-R*. American Psychiatric Association
- Watson J.C., Greenberg L. S. and Lietaer G. (1998). The experiential paradigm unfolding: Relationship & experiencing in therapy. In Greenberg L.S., Watson J.C. and Lietaer G., editors, *Handbook of experiential psychotherapy*, Guilford Press.