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

ISBN 978-88-3293-137-2

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
Cédrick 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 Tolouse, 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 Mingo: 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

Abstracts

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

Francesca Greco[1], Giulio de Felice[2], Omar Gelo[3]
[1]Sapienza University of Rome & Prisma S.r.l. – francesca.greco@uniroma1.it
[2] Sapienza University of Rome & NCU University – giulio.defelice@uniroma1.it
[3] University 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 ($\chi 2$, 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; 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 para-verbal 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	Passive	Familiar	Emotional		Daily
	(11.6%)	0.20	-0.56	-0.16	-0.01	-0.32
2	Transformative Process	Active	Unfamiliar		To empower	Daily
	(12.1%)	-0.46	0.29	0.06	-0.35	-0.16
3	Concrete thinking	Passive	Unfamiliar	Concrete	To empower	
	(16.1%)	0.84	0.34	0.42	-0.19	0.05
4	Therapeutic Relationship	Active	Familiar	Concrete	To understand	Relational
	(22.4%)	-0.25	-0.18	0.41	0.28	0.16
5	Relational Issues		Familiar	Emotional	To empower	Relational
	(14.6%)	0.04	-0.14	-0.47	-0.18	0.45
6	Feelings		Unfamiliar	Emotional	To understand	Daily
	(23.1%)	0.06	0.58	-0.43	0.49	-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 Textuelle*s, 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 Statistque 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 Firs 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.