The effects of lemmatization on textual analysis conducted with IRaMuTeQ: results in comparison

Mauro Sarrica[1], Isabella Mingo[1], Bruno Mazzara[1], Giovanna Leone[1]

[1]Department of Communication and Social Research, Sapienza University of Rome – Italy

Abstract

The software IraMuTeQ is gaining space in social and psychological research. It is free and easy to use, it provides quality outputs, and it fits with theoretical perspectives interested in communication and social construction of knowledge. As in other forms of automatized analysis of large textual corpora, its use involves pre-treatment and modification of the original text in order to reduce complexity. As we know, lemmatization is a very delicate phase, which affects the whole strategy of analysis (from the selection of lemmas according to statistical or substantive criteria, to the extraction of organising factors). However, algorithms implemented in commercial or free software, are often performed after the grammatical tagging, using reference dictionaries, in automatic and non-transparent way to the end-users. And, apart from anecdotical evidence, it is often difficult to evaluate the reliability of the automated procedures. The aim of this paper is to compare the outcomes of the procedures implemented by IraMuTeQ with the output obtained with other well established software. We used a large corpus in Italian language on the issue of "fiscal compact", consisting of over one million occurrences, drawn from over 3000 newspaper articles published from 2012 to 2015. The same corpus was lemmatised using the procedures available in IraMuTeQ (list based) and those implemented in Taltac2, Tlab, and Tree Tagger. The proximity between resulting lists of lemma produced by each software will be compared using intertextual distance. Finally, in order to examine the effects of different procedures on the textual analysis, we took the two most distant lists of lemmas and we applied a correspondence analysis to the two matrixes lemmas/newspapers.

Sommario

IraMuTeQ è un software che si sta diffondendo nella ricerca sociale e psicologica: è gratuito e facile da usare, offre risultati di qualità ed è adatto a prospettive teoriche interessanti nell'ambito degli studi sulla comunicazione e sulla costruzione sociale della conoscenza. Come in altre forme di analisi automatizzata di grandi corpora testuali, il suo uso comporta fasi di pretrattamento e di modifica del testo originale per ridurne la complessità. In questo processo, la lemmatizzazione costituisce una fase delicata per l'intera strategia di analisi: gli esiti della lemmatizzazione sono infatti propedeutici a quella di selezione dei lemmi, secondo criteri di rilevanza statistica o sostantiva, e a quella di estrazione di dimensioni testuali. Tuttavia, gli algoritmi implementati in software commerciali o gratuiti, sono spesso eseguiti dopo il tagging grammaticale, usando dizionari di riferimento, in modo automatico e non trasparente per gli utenti finali. A parte casi aneddotici, è spesso difficile valutare l'affidabilità delle procedure automatizzate. Lo scopo di questo lavoro è quello di confrontare i risultati delle procedure adottate da IraMuTeQ con quelli ottenuti mediante altri software più consolidati. A tal fine si è utilizzato un vasto corpus di lingua italiana sul tema del "fiscal compact", composto da oltre un milione di occorrenze, tratto da circa 3000 articoli di stampa pubblicati dal 2012 al 2015. Lo stesso corpus è stato lemmatizzato utilizzando le procedure disponibili in IraMuTeQ e quelle applicate da Taltac2, Tlab e Tree Tagger. La prossimità tra le liste di lemmi prodotte da ogni software è stata confrontata con la distanza intertestuale. Infine, per esaminare gli effetti delle diverse procedure sull'analisi testuale, si è applicata una Analisi delle Corrispondenze per le due matrici lemmi / testate più distanti.

Key words: Lemmatization, Software for Textual Analysis, Intertextual distance.

1. Introduction

IRaMuTeQ is a open software, distributed under license GNU GPL, based on R statistical software and on Python language[1]. It has now reached version 0.7 alpha 2 and it is still under development (Ratinaud, 2009).

IRaMuTeQ offers a range of treatments and tools to aid in the description and analysis of textual corpora and of people/characters matrixes, including analysis of specificities and similarity analysis (ADS). Probably, the feature that has received more interest is the implementation of the algorithm of descending hierarchical classification initially developed by Reinert and already implemented in Alceste (Ratinaud & Déjean, 2009; Ratinaud & Marchand, 2012a; Reinert 1983; Reinert, 1996). Moreover, IraMuTeQ has a user friendly interface and provides quality graphics output (that are even ameliorable thanks to the implementation wit Gephy[2]).

In sum, the characteristics of IRaMuTeQ make it a good option for researchers interested in conducting textual analysis on large textual corpora (Lahlou, 2012). Indeed, within the theoretical framework of social representations approach, this software has been used in association with other methods to examine various typologies of textual corpora, from political settlement speeches to twitter exchanges during mass emergencies (Sarrica, Germani and Brondi, 2014; Pola, Sarrica and Contarello, 2015).

However, as for any software, the usability of interfaces combined with the lack of knowledge of the details, exposes researchers to trust the outputs without questioning the way data are manipulated in each step by the software. In particular, attention often focus on the last steps of data analysis, whereas the first steps in which the corpus is prepared are given for granted. The cleaning of the corpus from non-alphanumeric tokens, that precede lemmatization, is a straightforward procedure which usually does not pose specific problems. More attention should be instead devoted to the morphological homogenization of the corpus and to its lemmatization. The perils of morphological homogenization have been examined in the case of contents spontaneously produced by people with different linguistic and technological knowledge (Ratinaud and Marchand, 2012b).

In this study we will examine the third step of text treatment: lemmatization.

The debate on the consequence of lemmatization on all the chain of analysis and on the quality of the final output goes beyond the scope of the present paper and will not be touched here (see Brunet, 2000; Mellet 2001, Tomasetto, Cattaneo, Selleri 2006). In any case, it is clear that, once opted for this procedure, the quality of lemmatization should be controlled. However, at best of our knowledge, the typology, lemmatization procedures implemented by IRaMuTeQ has not been systematically examined. In fact, lemmatization is often overlooked in research articles reporting the use of this software (Castro, Irene and Castanho, 2014; Sarrica, Germani and Brondi, 2014; Vivian et al., 2015). In other articles it is just described in a few words, for example: “First, the software replaces all terms by their canonical form

1 www.iramuteq.org; www.r-project.org; www.python.org
2 http://gephi.org

(lemmatization step): plural by singular, verbal forms by infinitive, elided words by corresponding non-elided words." (Delattre, Chanel, Livenais and Napoléone, 2015, p.65; see also in Pola, Sarrica and Contarello, 2015).

2. Aims

The current paper aims at providing a first answer to the following question: is the list of lemmas produced by IraMuTeQ comparable to those obtained by other software?

We were confronted with this problem while analysing a large corpus in Italian language. In order to provide a preliminary answer to our question we followed a comparative approach. Rather than proceeding to time-consuming manual screening of the entire list of lemma, we considered the list of lemma produced by different software as if they were different texts, and we examined to what extent they were rather near of far from one another (Labbé and Labbé, 2001). That is, we examined if the lists of lemma produced by different software were equivalent, partially overlapping or were so different to result being different corpora.

3. Method

In the next sections we will present the rationale we followed in the selection of the textual corpus analysed, the characteristics of the data, and the procedures adopted for their analysis.

3.1. Rationale for the selection of the corpus

The texts we are analysing in this study have been collected within a broader research conducted in collaboration with Fondazione Di Vittorio[3]. Goal of the research was to examine how the italian press described the *Fiscal Compact* (formally the Treaty on Stability, Coordination and Governance in the Economic and Monetary Union) and the introduction of the "balanced budget" principle into the text of the Constitution itself (law 1/12, 17 April 2012). By looking at how such a relevant change in the constitution had been presented to the italian citizens we aimed to highlight the level of maturity of democracy in Italy, and the underlying views of citizens participation. Through a careful choice of lexicon and contents, in fact, journals provide citizens with informed accounts, help to regulate emotions and to empower citizens, or on the contrary they foster fake participation which is based on managing consent and on the elicitation of the most regressive emotions.

In order to examine how the public debate on fiscal compact and on the constitutional change developed in time, we decided to collect articles published between January 2012 and May 2015. Moreover, in order to examine different voices, we collected articles published by seven media outlets. As a results of these two main objectives, we collected a vast amount of articles (see details below).

Text mining software are essential for managing such an amount of data. However, given the interdisciplinary nature of our research, we were confronted with the need to understand to what extent the results provided by different software and procedures could be confronted and triangulated. Proceeding backwards from the final results we identified lemmatization as the first step in which the use of different software may start producing significantly divergent outcomes.

[3] www.fondazionedivittorio.it

3.2. Characteristics of the corpus analysed

In this study we used a large corpus in Italian language on the issue of "fiscal compact", drawn from over 3000 newspaper articles published from 2012 to 2015 by seven italian newspapers and magazines. The exact dimensions of the corpus depend on the parsing and normalization procedures adopted by the software used: IraMuTeQ, Taltac2, Tree Tagger, Tree Tagger -Baroni list, Tlab. The softwares use different approaches to normalization and lemmatisation. Iramuteq, Taltac2 and Tlab use free context list-based procedures, in this study we used the lists implemented in the softwares, including multi-words detection during normalization phase. Tree Tagger is a probabilistic part of speech tagger that uses decision trees (Schmid, 2004), in this study we used the implemented parameters and the ones developed for Italian by Baroni.

As a result N varies from 1017421 to 123748 word-tokens; V varies from 40000 to 45895 word types; after lemmatization L varies from 24232 to 34589 lemma types (see Tab.1).

3.3. Procedure of analysis

In order to reduce the non-significant variation of the corpus, the analyses were conducted on two typologies of transformed corpus:

(1) the Transformed Corpus (TC) includes all the lemmas obtained through the reduction of capital letters in lower case and after the elimination of numerical forms. The TC thus includes all available information;

(2) the Transformed Corpus without Stop words (TCwSW) includes a subset of lemmas obtained by removing some stop words (SW). The TCwSW provides a focus on the more meaningful entries, consistently with content analysis applications.

The characteristic of the Transformed Corpus (TC) and of the Transformed Corpus without Stop words (TCwSW) are displayed in tab.1 .

	IraMuTeQ	Taltac2	Tree tagger	Tree tagger Baroni	T-lab
			Corpus		
Tokens (N)	1062828	1017421	1237481	1237481	1129566
Forms (V)	41890	45895	45577	45577	41655
Lemmas (L)	24650	34589	24232	25348	32962
			Transformed Corpus (TC)		
Tokens (N)	1045917	999147	1083231	1082533	1117788
Lemmas (L)	23885	31685	23118	23669	31903
			Transformed Corpus without Stop words (TCwSW)		
Tokens(N)	597186	581157	617119	642151	659916
Lemmas (L)	23815	31611	23088	23567	23892

Table 1. *Characteristics of the corpus by software*

Following a comparative approach, we examined the outcomes of the lemmatization procedures conducted on TC and on TCwSW using IraMuTeQ with the outputs obtained with the following other software: Taltac2, Tree Tagger, Tree Tagger -Baroni list, Tlab[4].

The dissimilarities between each pair of the lists of lemmas drawn from different software were calculated using intertextual distance (ID), developed for stylebased classification (Labbé and Labbé, 2001; Labbé, 2007). It is based on a sum of differences between the frequencies of words (lemmas in our case) in two texts (lists of lemma in our case). Given a pair of texts A and B of size N_A and N_B with $N_A \leq N_B$, their ID is:

$$ID_{(A,B)} = \frac{\sum_{i \epsilon V_{A \cup B}} |F_{ia} - \hat{F}_{ib}|}{N_a + N'_b}$$

where $N'_B = \sum_{i \epsilon B} \hat{F}_{ib}$ and $\hat{F}_{ib} = F_{ib} \times \frac{N_A}{N_B}$

$\hat{F}_{ib}$ is the frequency of each i type in B reducing according to the size of the smallest text (A). If the lengths of the two texts are very different, the calculation of ID should be limited to the B types whose frequencies $\hat{F}_{ib} \geq 1$ (Labbé and Labbé, 2001, Labbé, 2007). In our case, the application of two procedures leads to similar results.

According to the ID standardized scale, the value 0.25 is the threshold for considering two texts having the same characteristics in terms of author, genre or topic (Labbè and Labbé, 2001 p.8-9). In our study the only source of variation is the software used, so values higher than this threshold would indicate that the output lists are quite different and that the lemmatization procedures implemented by the software create different corpora.

Finally, in order to examine the effects of different procedures on the chain of analysis, we took the two most distant lists of lemmas and we applied a correspondence analysis to the two matrixes lemmas/newspapers. The lemmas with frequency >30 were included (IraMuTeQ N=501056; Taltac2 N= 454317), which cover above the 75% of TCnSW and the 45% of TC.

4. Results

The intertextual distances between the lists of lemmas are showed in table 2. The lemmatization applied to TC produced different lists (ID1 values >.25) in seven out of ten comparisons, even though some of the values are very close to the threshold. Conversely, as regards the lemmas extracted from TCwSW, six comparisons show that the lists are similar (ID2 ≤ .25).

The ID between lemmas identified by IraMuTeQ and by Tree Tagger is good, it is acceptable as regards TLAB and non-acceptable as regards the output obtained by Taltac2, which utilizes normalization procedures based on specific lists for Italian language.

The two matrixes lemmas_ IraMuTeQ /Newspapers and lemmas_Taltac2/Newspapers were submitted to correspondence analysis in order to examine the effects of lemmatization on the textual analysis. Despite the ID between the output produced by the software, the results of

[4] Taltac2 v.2.10 (www.taltac.it); Tree Tagger (http://cis.uni-muenchen.de); Tlab 8.12 (http://tlab.it/it/)

the correspondence analysis are similar. Both analyses identify six dimensions, with similar eigenvalues (see Tab. 3 for the first three dimensions).

	ID1- Transformed Corpus (TC)				ID2 - Transformed Corpus without Stop words (TCwSW)			
	Tree Tagger	Tree tagger Baroni	Taltac2	Tlab	Tree Tagger	Tree tagger Baroni	Taltac2	Tlab
IraMuTeQ	**0,1166**	**0,1806**	**0,3032**	**0,2631**	**0,1374**	**0,1753**	**0,2647**	**0,2506**
Tree tagger	0,0000	0,0965	0,3224	0,2620	0,0000	0,0623	0,2823	0,2409
Tree tagger Baroni		0,0000	0,2975	0,2911		0,0000	0,2995	0,2466
Taltac2			0,0000	0,2578			0,0000	0,2988

Table 2. *Intertextual distances*

Moreover, though the plots are symmetrical (Fig.1 and Fig. 2)[5], the results lead to the same interpretations: for both analyses, dimensions 1 and 2 oppose technical vs. political language, and international vs. local politics respectively.

	IraMuTeQ			Taltac2		
Dimension	Eigenvalue	Percentage	Cum Percentage	Eigenvalue	Percentage	Cum Percentage
1	0,0251	30,71	30,71	0,0285	29,35	29,35
2	0,0140	17,06	47,77	0,0167	17,22	46,57
3	0,0137	16,72	64,49	0,0162	16,68	63,25

Table 3. *Correspondence Factor Analysis – Eigenvalues of the first three dimensions*

Also the positioning of newspapers is the same (Fig. 3 and 4): Il Sole 24 Ore is linked with technical language, Il Fatto Quotidiano focuses on national politics, whereas the Il Giornale depicts international politics with emotionally loaded terms.

[5] All active lemmas were considered in the interpretation of the dimensions. However, in order to make the comparison easier, only the 15% of lemmas with the highest contributions are represented in the figures 1 and 2.

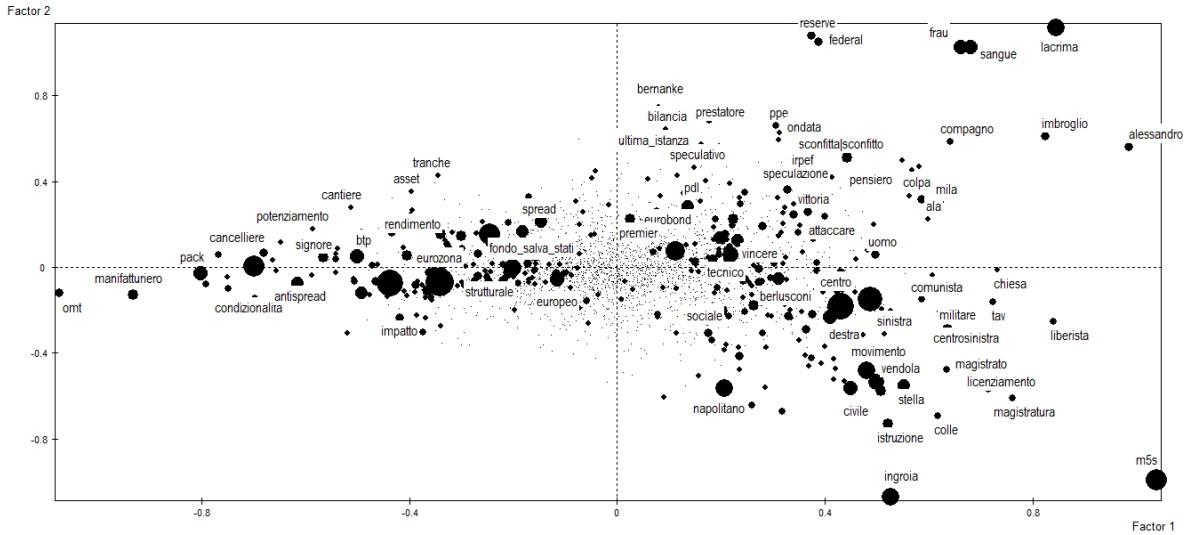

Figure 1. *IraMuTeQ - Correspondence Analysis – Plot of active cases (lemmas) on dimensions 1 and 2*

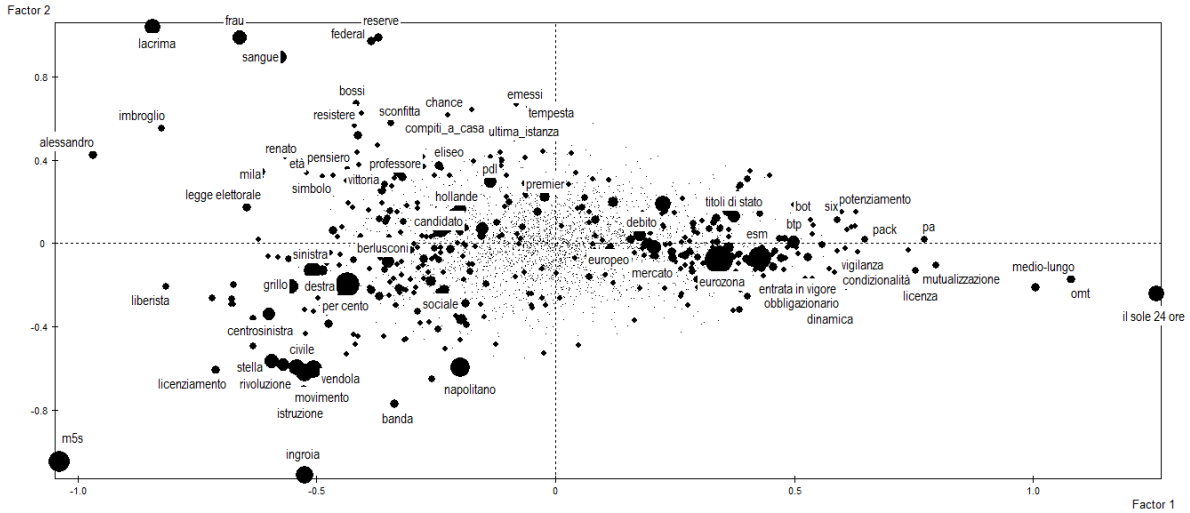

Figure 2. *Taltac2 - Correspondence Analysis – Plot of active cases (lemmas) on dimensions 1 and 2*

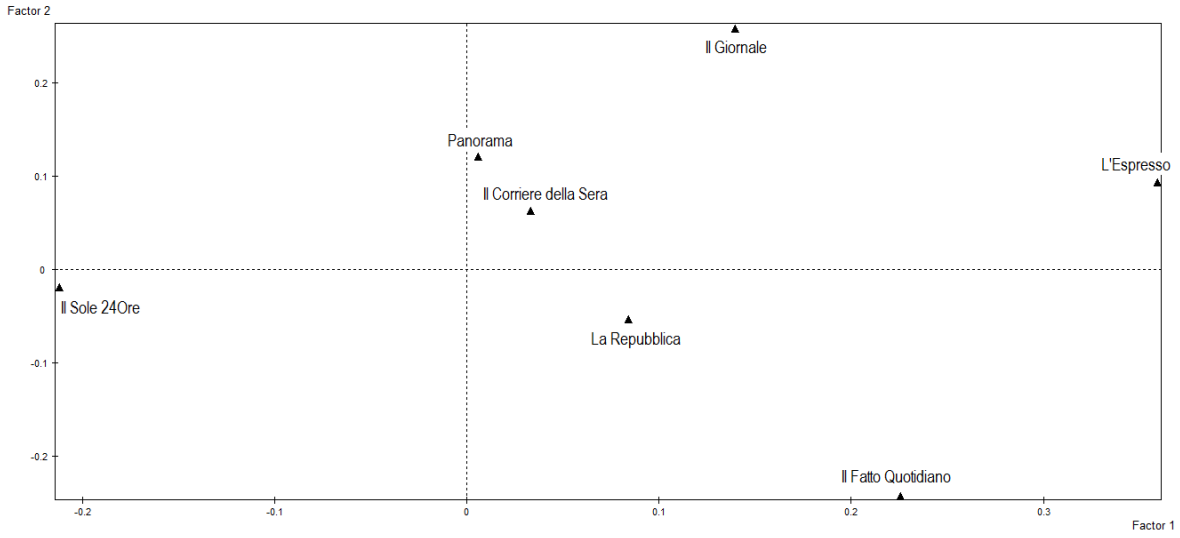

Figure 3. *IraMuTeQ - Correspondence Analysis – Plot of active frequencies (newspapers) on dimensions 1 and 2*

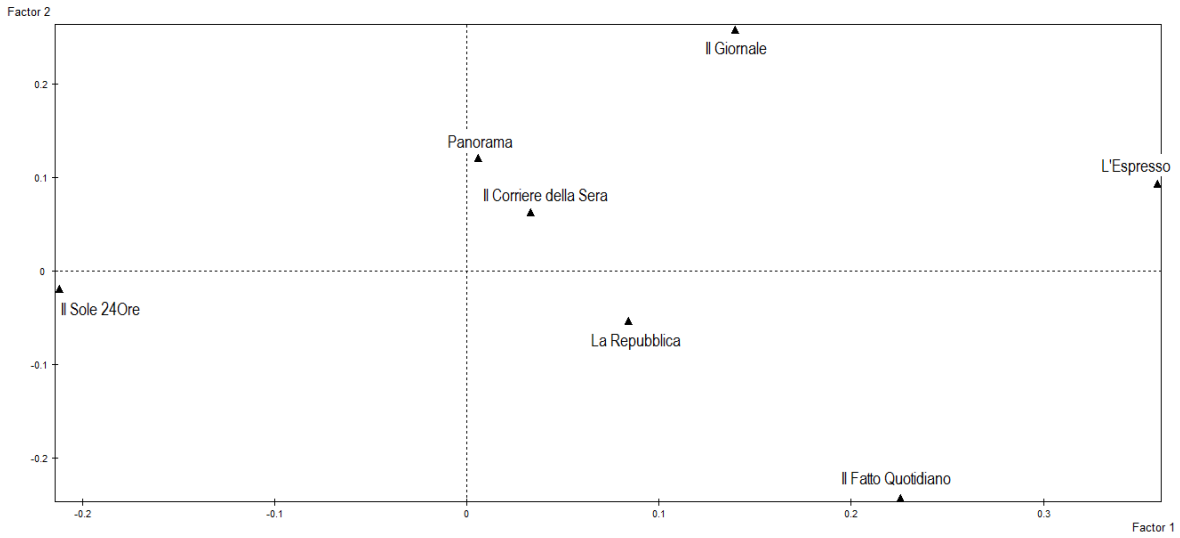

Figure 4. *Taltac2 - Correspondence Analysis – Plot of active frequencies (newspapers) on dimensions 1 and 2*

5. Conclusions

In this contribution we provided preliminary tests of the lists of lemmas produced by IraMuTeQ. In fact, the usability of IraMuTeQ and the procedures it provides make it an interesting options for conducting textual analysis on large textual corpora (Lahlou, 2012; Sarrica et al., 2015). However, while using this software, we were confronted with the problem of showing that the lemmatization produced by IraMuTeQ is comparable with those obtained by other well-known software. Our results provide a positive answer to this question.

First, the analysis of the ID shows that the lists of lemma can be considered the same or very similar in most of the cases. It should be noted that we intervened only to a minimal extent in

cleaning the corpus. Nevertheless, especially as regards the Transformed Corpus without Stop words (TCwSW), all the distances are acceptable with the exception of those with Taltac2.

Further investigations are needed to understand this result. However, a preliminary explanation can be given. Taltac2, in fact, is especially suitable for the Italian language, also for its capacity to recognise multiwords and names. IraMuTeQ, even though it has a dictionary of expression for Italian, doesn't have this level of specificity. This could be proposed as a refinement for further development of the project. Despite such a difference, however, the results of CA conducted using the two most distant lists are very similar. In sum, the final interpretation of the analysis is comparable. These results can be interpreted in general terms as an evidence of the relative irrelevance of lemmatization procedures on large corpora (Brunet, 2000; Mellet 2001, Tomasetto, Cattaneo, Selleri 2006). However, as regards the goals of this contribution, it further show the solidity of the output reachable by using the new software.

References

Brunet, E. (2000). "Qui lemmatise dilemme attise". *Lexicometrica*, 2 : 1-19. http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.PDF

Castro, N. Q., Irene, M. and Castanho, S. (2014). Análise crítica do debate presidencial das eleições de 2014. *Psicologia E Saber Social*, *3*(2): 260-266.

Delattre, L., Chanel, O., Livenais, C., and Napoléone, C. (2015). Combining discourse analyses to enrich theory: The case of local land-use policies in South Eastern France. *Ecological Economics*, *113*: 60–75.

Labbé, C. and Labbé, D. (2001). Inter-textual distance and authorship attribution. Corneille and Moliere. *Journal of Quantitative Linguistics*, *8*(3): 213-231.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, *14*(1): 33-80.

Lahlou, S. (2012). Text Mining Methods: An answer to Chartier and Meunier. *Papers on Social Representations*, *20*, 38.1–38.7.

Mellet, S. (2001). Lemmatisation et encodage grammatical: un luxe inutile?. *Lexicometrica, 3-http://lexicometrica.univ-paris3.fr/article/numero3.htm*

Pola, L. G., Sarrica, M. and Contarello, A. (2015). Imprenditori di identità a Palazzo Marino. Cittadinanza e valori nei discorsi di insediamento dei Sindaci di Milano dal dopoguerra a oggi. *Psicologia Sociale*, *3*: 223–256.

Ratinaud, P. (2009). *IRAMUTEQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. http://www.iramuteq.org.

Ratinaud P. and Déjean S. (2009). RaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre, *Modélisation Appliquée aux Sciences Humaines et Sociales*, 8-9.

Ratinaud, P. and Marchand, P. (2012a). Application de la méthode ALCESTE aux «gros» corpus et stabilité des «mondes lexicaux»: analyse du «CableGate» avec IRAMUTEQ. *Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles*, 835-844.

Ratinaud, P. and Marchand, P. (2012b). Recherche improbable d'une homogène diversité : le débat sur l'identité nationale. *Langages*, *187*(3): 93-107.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse Des Données*, *8*, 187–198.

Reinert, M. (1996). Un logiciel d'analyse lexicale: ALCESTE. *Les Cahiers de L'analyse Des Données*, *XI*, 471–484.

Sarrica, M., Germani, M., and Brondi, S. (2014). Istanti, ore, giorni dopo il terremoto. Spunti dall'analisi dei tweet per la Psicologia dell'Emergenza. In Comunello, F., editor, *Emergenze "social": Il ruolo dei social media nella comunicazione d'emergenza*, pages 87-106. Roma: Guerini Editore.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the international conference on new methods in language processing,* 12, 44-49.

Tomasetto, C., Cattaneo, C., Sellerio, P. (2006). Molto rumore per nulla? Gli effetti della lemmatizzazione sull'analisi di un corpus di interviste con ALCESTE. *Act Jadt2006*, 907-919. http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-081.pdf-

Vivian, R., Brangier, E., Barcenilla, J. Bornet, C., Roussel, B., & Bost, A. (2015). Towards Participatory Methods to Take into Account Future Users and Future Usages of Hydrogen Energy: a Prospective Ergonomics Approach. The Eighth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services. *CENTRIC 2015. November 15-20, 2015 - Barcelona, Spain.*