Text Mining on Elementary Forms in Complex Lexical Structures

Simona Balbi [1], Sergio Bolasco [2], Rosanna Verde [3]

[1] Dip. di Matematica e Statistica – Università "Federico II" di Napoli – Italia – sb@unina.it

[2] Dip. Studi Geoeconomici, Linguistici, Statistici SAR – Università "La Sapienza" di Roma – Italia – sergio.bolasco@uniroma1.it

[3] Dip. di Strategie Aziendali e Metodologie Quantitative – Seconda Università di Napoli – Italia – rosanna.verde@unina2.it

Abstract

After showing the advantages of formulating lexical structures with variable elements in terms of symbolic objects (*in this issue*), the Authors propose to introduce the *a priori* information which determine their building in the analysis of elementary *textual* units (*extensions*). It is worth noting that, dealing with symbolic data, the observed textual units disappear by the collapsing procedure. In order to visualize forms, an analysis on elementary data, introducing the external information on the complex structure they belong, has been proposed. This analysis can be usefully performed complementary to the symbolic objects analysis, because it enables to analyze the dependence relations of the forms on the contextual information in which they have been used. In order to enrich the analysis of textual data, it is possible to introduce other external information, related to the fragments where the forms appear. In doing that by a double partial analysis, we represent on low-dimensional spaces the relational structure existing between the two sets of information introduced. Forms and fragments can be represented as supplementary points, in order to study the role they played in those relations. An application dealing with a very large *corpus* of lexical structures with variable elements, extracted from the Italian newspaper "La Repubblica" during the Nineties, has been performed, in order to show the relation between years and contextual information in the different identified context, and the single forms mainly involved.

Keywords: corpus linguistics, text mining, symbolic data, factorial analysis, projection operators.

1. Introduction

When we aim at an effective description of a long text, dealing with "automatic reading" tools (and not by a "direct" reading), it is today preferred to value the whole content of the text together with other overall elements. These elements must be able to reproduce what we can call the "imprinting" of the text, that is to say, a *pattern* of standard characteristics able to form a sort of DNA.

The present work can be included in the theoretical framework of the so-called *corpus linguistics* (Biber *et al.* 1998, Habert *et al.*, 1997). Aiming at proposing new text mining procedures, the paper bases its statistical analysis on a very large database formed by more than 250 million occurrences which appeared in an Italian newspaper. Starting from this huge database, we have tried to give elements profitable in understanding the previously defined *pattern*. On the one hand, we have introduced some general references on the basic characteristics of the text; on the other hand, we have experimented the visualisation of some complex information with a new graphical technique.

In the following, in particular, once some text "imprinting" characteristics have been defined, their impact in the reference database have been measured. After recalling a text mining procedure concerning the identification of complex lexical structures and their formalisation in symbolic objects terms (Bolasco, *et al, in this issue*), a non symmetrical correspondence analysis (Lauro and D'Ambra, 1984) has been performed in order to achieve information on the elementary forms belonging to the complex structures.

2. Corpus and related database

The source that is used to measure some of the identified characteristics are 10 years of the newspaper "La Repubblica". The *corpus* (Rep-90) has been obtained from the Cd-Roms containing all the articles appeared in the newspaper each year. In order to understand how large the texts are, we have to think that, the articles published in a year produce more than 20 million occurrences and almost 300.000 different graphical forms[1].

The text analysis has been done for each year. A subsequent process has combined each different vocabulary in an annual occurrence matrix for a temporal evolution analysis. Distinguishing forms of capital letters from small letters, various databases have been built. Amid the several aims of a collection like this, there is the purpose to compile an inventory of a large range of words with relating prefixes and suffixes in order to test various tools and data/text mining procedures.

For our work, we have activated automatic procedures in order to draw out words starting from graphical forms, leaving out "no words" (that means spelling mistakes, abbreviations, numbers) and nouns (proper nouns and toponyms). Just considering frequent famous people nouns and toponyms, we have about the 10% of the occurrences, with a very large variety of forms (dozens of thousands).

We have first prepared a main reference database that adds up more than 221 millions of occurrences and 280.325 words with small letters[2]. In table 1 we have a first estimate of the situation with reference to forms without nouns, numbers, abbreviations and spelling mistakes.

Of each word we know the occurrences for years, grammatical category (if it is one) and the origin of the word capturing process. This is very important (considering the database extension) because many graphical forms are, in fact, derivatives of a basis[3] (e.g. *person*)

[1] We obtain such a large number of forms considering capital/small letters. In newspapers, nouns and abbreviations are very numerous. Therefore the number of words in a dictionary is lower than graphical forms number (also including foreign words). Moreover (as showed by Silberzstein (1995) in a work on two years of "Le Monde"), in newspapers, graphical forms, as simple chains of letters, not always correspond to words very often because of spelling mistakes.

[2] Among those words we have also considered the most important hyphenated words (e. g. baby-sitter, vetero-comunista), useful in order to identify compounds. Successively, we have standardised forms, by eliminating the hyphen and, in the case of double vowels <*antiitaliano*>, <*neooperatore*>, <*ultraavanguardie*>, we have collapsed the forms in order to add the occurrences of <*anti-italiano*> with those of <*antiitaliano*> and <*antitaliano*>. However, more than 180.000 words with small letters are still unclassified.

[3] In newspapers some derivative forms related to famous people nouns exist (e.g., *neodarwinismo*, *anticraxista*, *nannimorettiani*, etc.).

obtained using prefixes and suffixes (e. g. *person*/ale *person*|alismo, *person*|alizzazione, im|*person*|are, pesa|*person*|e …).

TAB. 1 – Importance of small letter forms (without nouns) identified in Rep-90 corpus

Original Tagging	Graphical forms		Occurrences	
	a. v.	%	a. v.	%
Words	179,652	64.1	212,489,462	96.0
Suffixes	52,775	18.8	954,692	0.4
Prefixes	35,878	12.8	309,316	0.1
Various forms	12,020	4.3	7,525,746	3.4
Total	280,325	100.0	221,279,216	100.0
Legenda: "various forms" = compounds, foreign words, apocopated words, etc.				

As known, dictionaries do not have all the variants of a word and we need *ad hoc* procedures[4] for identifying them. In tab. 1 we can note an increasing of almost 30% of forms identified by suffixes and prefixes not found in the dictionary[5]. As we can observe, they usually are low frequency forms because, on a whole, they have a 0,5% impact in occurrences terms[6]. Therefore, if we want to do complete text-mining operations, not only do we have to consider the flexion base of a lemma, but we also have to consider prefixes and suffixes obtained by derivation.

3. Some characteristics of the "imprinting" of a text

The elements able to characterise the "imprinting" of a text can vary a lot, depending on the aims. Considering the content of the text it is however possible to define some frequent indicators able to describe each text. We should think about the punctuation marks, the use of difficult words (in first approximation, considering the word length), the length of the sentences, the different parts of the speech structure (e. g. macro-grammatical category proportions: N, A, V). We should also think about the distribution of the so-called "instrumental words", which are lacking in content but essentials – as a whole - in order to individualise some pillars in the speech structure (we have to think about the differences in written and spoken language – Biber *et al.* 1998, p.106, Halliday 1994, Voghera 1992).

Consequently it is not difficult to draw statistical indicators from the Rep-90 *corpus* in terms of normalised occurrences per million words. This very large *corpus* could be a good reference for the Italian standard contemporary language (*see* Tab 2).

Moreover, studying these characteristics means collapsing too large matrices of textual data in order to obtain useful syntheses able to reconstruct the pattern of the searched imprinting.

[4] In Taltac software (Bolasco, 2000) we have some queries able to capture those variants. Therefore, it is possible to extend the coverage of the reference dictionary used for the grammatical tagging.

[5] At the present moment, we have considered the most fruitful suffixations (*-mento*, *-zione*, etc.) and a set of about 350 most common prefixes. The work in progress has to find, among the 180000 unidentified forms, the ones that can be still extracted.

[6] From Rep-90 corpus, it is possible to obtain some quantifications of this phenomenon: for example, the verb flexions, that usually in Italian are less than fifty, can be 150-180 for the most common verbs, when we consider pronominal reference, aiming to point out the relation between a subject and an object.

Tab. 2 - Some characteristic of the Rep-90 corpus

Puntuaction	Occ. x million words	%
,	66,904	50.3
.	46,786	35.2
:	7,097	5.3
)	4,191	3.1
(	4,038	3.0
?	2,547	1.9
;	1,331	1.0
!	170	0.1
Total	133,065	100.0

CAT	% types	% tokens
A	15.5	6.3
AVV	1.9	4.3
CONG	0.0	0.8
DET	0.0	6.3
ESC	0.0	0.0
N	21.7	42.6
PREP	0.0	7.0
PRON	0.1	3.8
V	60.9	28.8
Tot. not ambiguous	100.0	100.0

Length	% types	% tokens
1	0.01	7.524
2	0.06	19.303
3	0.29	14.023
4	0.93	7.157
5	2.08	12.385
6	3.83	9.334
7	7.81	8.307
8	10.38	7.376
9	13.60	5.159
10	14.97	4.008
11	13.72	2.490
12	11.35	1.283
13	8.10	0.819
14	5.54	0.491
15	3.29	0.216
16	1.95	0.072
17	1.04	0.035
18	0.53	0.014
19	0.27	0.004
20	0.13	0.002
21-22	0.10	0.001
23 e oltre	0.04	0.000

In addition to these general elements, it is possible to define more elements of general interest that permit to synthesise the text according to some criteria, which are useful for expressing quick judgements on the speech content (e. g. the characterisation of some important grammatical categories).

For example, if we consider the verbs, defining the whole structure of the types/tokens regarding *past*, *present* and *future* tenses is an element that depicts the text. Also the study of the subjects (I, you, he-she-it, we, you, they), obtained from verbal flexion or a pronominal reference[7], is useful to identify the main persons in the discourse.

In general, the word-formation process (e. g. nouns or adjectives derived from a verb) helps us to point out some groups of terms useful for our aims. Considering the adjectives, we can focus our attention on those ending in *–evole* ("capability" with active and passive meaning) or in *–bile* (passive meaning of possibility)[8]. For nouns, we have e. g. abstract nouns ending in *-ezza*, *-tà*, *-itudine*, *-aggine*, *a/enza*, *ismo*, *-esimo*, etc. or nouns meaning an action, ending in *-mento*, *-zione*, *-sione*, *-tura*, *-aggio*, or a person/thing doing an action such as the words ending in *-s/tore*, *-trice*, *-a/ente*, *-iere*, *-ista* etc. Those classes are not homogeneous and exclusive in relation to these characteristics (Bolasco *et al.*, *in this issue*), therefore the researcher supervision is always necessary.

[7] Very useful to point out the relation among a subject and an object (*buttameli*) or in order to specify references to person in some tenses, as the gerund (for example, *facendomi/ti/vi*).

[8] But also in: *-a/ente, -ivo, -ato, -uto, -are, -ario, -ale, -ano, -aceo, -a/ineo, -igno, -ino, -izio, -iero, -esco, -ico, -a/istico, -ifico, -torio, -oso*: Or the superlatives in *–issimo*, or the numerals.

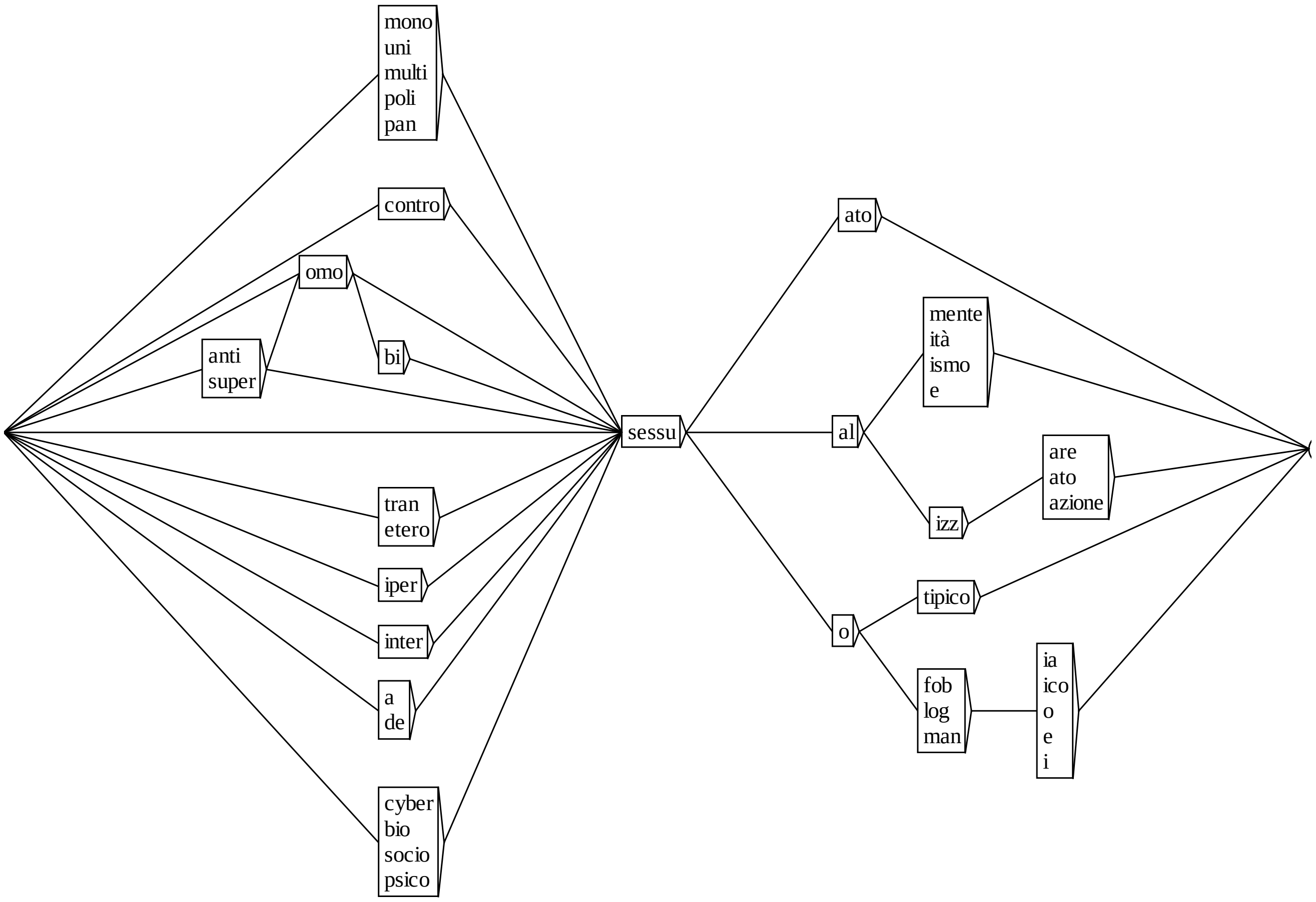
4. Definition of complex structures

However, interesting elements mining is often based on the definition of more complex lexical structures. Observing in a text, for example, every word related to the headwords <grafic> or <politic>, our interest is focused on all their derivatives forms and on all their developments considering their prefixes. From our experimental studies, we have observed that identifying the prefixes allows us to significantly increase the number of recognised and classified forms (more than 12% of the vocabulary, *see* Tab. 1).

In fig.1 we illustrate the graph derived from the query "*sessu*" in Rep-90 Corpus, which is the starting-point of all the topic "sex" derivations.

By analysing the graph, which is reconstructed on a Base, we are able: to sum up all the tokens of the studied topic and rendering less ambivalent various possible usages of the term combining suffixes and prefixes.

*Fig. 1 - Morphological graph derived from the query *sessu* in Rep-90 Corpus*



If we want to know which other Bases are similar to **sessu** in standard language we have to consider the similarity of their derivative and prefixes graphs in Rep-90 corpus. We can observe that, among the most frequent words, the root of the word "*persona*" (<person>) is very much correlated in regards to the suffixes and it is less correlated in regards to the prefixes (showing a different distribution). This with an exception of the branch in <o> and its derivative forms (e. g. [[*fob*]*ia/ico*]). On the contrary, the graph of the word <politic>, more productive of prefixes than in suffixes, is much more correlated with <economic> than with <person>and so on[9].

What we have showed up to now allows us to study specific areas of meaning. However, we have to consider the problem of having a very large quantity of information that we must use suitably for the quantitative analyses.

We can't actually limit our study collapsing the original matrix by summing up all the forms related to an area of meaning. On the contrary, we propose to use a formalization of the textual data, in statistical terms, that are able to consider the complex structure in these analysed forms subsets.

In another work (see Bolasco *et al.*, *cit.*) this has been widely illustrated. In the following paragraph it is briefly described but in par. 6 and 7 there is a different application of it, using the same model of complex structure.

This structure derives from 2428 elementary forms present in Rep 90 originated by *-ismo, -ista, -istico, -iano* suffixes associated to specific roots concerning geopolitics topics with the anti/ex/filo/post/ultra prefixes strongly related to the dealt topic.

The selected bases have been assembled in 10 groups and those groups have given rise to 10 symbolic objects (Bolasco *et al.*, *cit.*). From this sub-database we have developed the statistical analyses shown in par. 6. Let's now explain the symbolic data concept.

5. The symbolic data formalisation

Diday proposed Symbolic Object as a new tool for modelling complex data as well as *concepts,* deriving from "*real world". Symbolic Objects (SO's) allow to overcoming the classical definition of statistical units characterised by single values for each observed variable, whereas SO's are described by multi-valued variables and they keep in their description the relational structure inside complex data (e.g.: hierarchical structure, logical constraints).

SO's can be considered as generalised data which usually arise by: expert knowledge (e.g.: species of plants; scenario of accidents - or in textual context: *grammatical categories, verbal forms*); classification of individuals in homogeneous classes having a conceptual meaning (e.g. *variable graphics forms*). Moreover, in knowledge extraction framework, SO's allow one to synthesise huge sets of data, stored in databases, in terms of underlying *concepts* (e.g. *contextual patterns,* in textual mining). A description of these concepts is expressed by structured data (so called, *symbolic data*) characterized by sets of values that they take with respect to more complex type of variables (*symbolic variables*), which allow to taking into account their internal variability (e.g. *number of the variants in a graphic form*) as well as by relationships defined among these descriptors.

Finally, in a recent book on *Symbolic Data Analysis* (Bock, Diday, 2000), a *symbolic object s* is defined by the *triple (a, R, d), where "d" is a description* - given by the sets of values taken by the *symbolic* variables y_1, …, y_p (multi-categorical, interval; modal) - *"R" is a relation between descriptions* - defined by a sub-set of operators $\{=, \in, \subseteq, \dots\}$- *and "a" is a*

[9] Are very fruitful, usually, queries on the most important word of a corpus. Good examples have been done on Rep-90 corpus for this bases: *person, politic, econom, cultur, sistem, privat, pubblic, capital, special, grafic.*

mapping function which allows to compare the descriptions of a set of elements ω_i in Ω to the description of the object *s*.

Let *d'* be a description of an element to be compared with *d* of *s* by *R* (the operator to compare the two description is chosen according to the kind of descriptor): [d'Rd] $\in L$ where *L*={*true, false*}. Thus, "*a*" is a recognition or allocation function which allows to assign each ω_i to the extension of the object s if *L=true*. The extension of an object is given by:

$$\text{Ext}(s|\Omega)=\{\omega\in\Omega|\ a(\omega)=[\text{d'Rd}]=\textit{true}\}$$

This way to compute the extension of a *textual* symbolic object can be considered as an useful tool to recognise unknown forms as belonging to a *textual category* (defined as a *textual SO*).

6. Analyses of elementary data and complex information

The aim of the analysis on elementary data is to obtain a visualisation of the single graphical forms with respect to the structure of relations existing among the elements defining the complex data and their presence in standard fragments.

As known, lexical correspondence analysis (Lebart *et al.*, 1998) is generally used to identify the principal components of the association structure in a lexical table **T**. **T** cross-classifies forms and fragments (sub-texts), often aggregated according to a partition variable. In this case, Balbi (1995) proposed to study the vocabulary dependence on the partition variable, through a non-symmetrical correspondence analysis (Lauro, D'Ambra, 1984). In this way, we introduce in the analysis elements of knowledge that are outside the contingency table.

The idea of introducing external information related to both lexical table dimensions is fruitful. It allows considering in the analysis information not only on sub-texts but also on forms. It is so possible to introduce information on context or on corpus pre-processing.

In this work, in particular, we refer to Balbi and Giordano (2000) proposal of a correspondence analysis on a table that includes information about the analysed *corpus* (both on forms and fragments), by means of a doubly partial analysis. It is mainly a geometrical approach that considers the study of dependence in terms of projections on the subspace, generated by the matrix elements that contain the external information.

The most important purpose of this work is to recover the existing relation among an analysis on the data derived from the formal definition of a concept (intention) and its composition in terms of elementary units (extension). In analysing complex data structures, it is therefore important to introduce a collateral analysis able to preserve the relations among the single elements and their related object. Dealing with textual data, this is useful to connect graphical forms and complex lexical structures.

If (as in our application) the information on fragments is related to time, it is possible to analyse the temporal evolution of the phenomenon, also by graphical representations (for example, trajectories drawn by linking points representing a specific word used in different times and contexts).

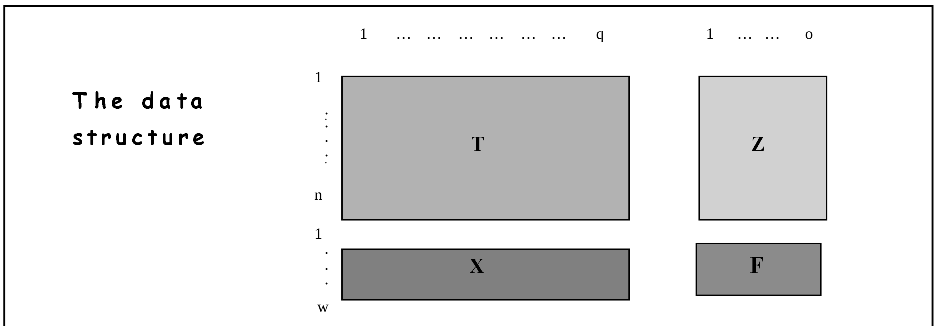
In the following, for the sake of simplicity, we show this method referring to the lexical structures with variable elements, previously introduced, observed in Rep-90 corpus. We consider the original corpus as formed by all the articles published during the Nineties and the aggregation criterion is given by years of publication. Therefore, let's consider the following matrices:

T (*forms, articles*) is the lexical table (*n*, *q*), **X** (*years, articles*) is the indicator matrix (*w*, *q*) classifying the articles and **Z** (*forms, objects*) with dimensions (*n, o*) is the indicator matrix that adds graphical forms belonging to the same symbolic object.

It is very important to consider the graphical representations given by this geometrical approach. These graphics are based on projection operators on subspaces generated by the reference matrices **X** and **Z**. In particular, this analysis aims at understanding the use of the considered lexical structures, in relation to the year when the article was written. Moreover, it is possible to represent, as supplementary points, the single graphical forms, underlining, from the first factorial plane lecture, the role played by some forms in the interpretation of the factorial pattern. Similarly, it will be possible to project, as supplementary points, aggregated forms, obtained by the union of single forms (e.g. those where there is the prefix *anti-*, or those with the suffix *–ismo*, or the base *marx*).

Let **P** be the (*o*, *q*) matrix obtained by the product **Z'T**, with general element $p_{s,j}$, frequency of the *s*-th object in the *j*-th article. By considering the row and the column marginals of **P**, we define the diagonal matrices $_r\mathbf{D}_p$ and $_c\mathbf{D}_p$, containing respectively the **P** row- and column-marginal distributions. Let us name **F** the (*o*, *q*) column profile matrix given by $\mathbf{P}_c\mathbf{D}_p^{-1}$. Its general element is the conditional frequency of objects on articles (*see* Fig.2). In order to introduce information on sub-texts (articles), we consider the effect due to the aggregation of article according to the information in **X** (year of publication). From a geometrical viewpoint, we aim at analysing the orthogonal projection of **F** on the subspace generated by the rows of **X**. Therefore, we study the dependence of the expert information (which lead us to build the previously defined symbolic objects) on the information on articles (year of publication). Thus, the orthogonal operator is $\mathbf{X'(XX')^{-}X}$ and the analysed matrix $\hat{\mathbf{F}} = \mathbf{FX'(XX')^{-}X}$.

Fig. 2 – The relations among the matrices in the non symmetrical analysis



T is the lexical table with general element t_{ij} , the frequency of the *i* form in the *j*-th article

X is the indicator matrix with the external information on articles, i.e. the year when the article appeared

Z is the indicator matrix with the external information on forms: the *o* columns indicate the presence (1) or absence (0) of a form in an object

F is the matrix cross-classifying articles×objects. On **F** the non symmetrical analysis is performed

The procedure we propose consists in performing a non symmetrical correspondence analysis on $\hat{\mathbf{F}}$, by means of a generalised singular value decomposition of the profile matrix $\mathbf{D}_o^{-1}\hat{\mathbf{F}}$. The orthonormalising constraints are $\mathbf{u}'_\alpha\mathbf{D}_o\mathbf{u}_\alpha = 1$ and $\mathbf{v}'_\alpha\mathbf{v}_\alpha = 1$, where $\mathbf{D}_o$ is the diagonal matrix (*o*, *o*) of the symbolic objects marginal distribution. As a consequence, $\mathbf{D}_o^{-1}\hat{\mathbf{F}}$ has in columns the conditional distributions of the criteria defining the symbolic objects, with respect to the years of publication in "La Repubblica".

From a geometrical point of view, the different constraints lead to a different way to evaluate distances between points: distances between *years* are calculated in an Euclidean weighted metrics, while the distances between the objects are calculated in a usual Euclidean metrics.

The simultaneous representation (*joint plot*) of the row-points and the column-points must consider the different metrics in both spaces. In fact, we can measure the distance only between points that are in the same space. Instead, we can evaluate the position of each point in a space only in relation with the whole cloud of points related to the other variable categories (for additional details on how to read factorial planes in non symmetrical analysis, *see* Balbi, 1997).

In order to represent the forms as supplementary elements, we have to consider the row marginal frequencies of **Z** and to build the matrix of the row profiles $\mathbf{R} = {}_r\mathbf{D}_z\mathbf{Z}$.

After centring, profiles can be projected on the factorial planes obtained by the previous analysis.

Summarising, the proposed analysis consists in transforming the whole amount of elementary data forming the original lexical table **T** (containing all the forms appeared in each article) in a new table cross-classifying expert information (i.e. the objects) with the years of publication.

A generalised singular value decomposition (with constraints related to the hypothesised relation between the two ways of the table) supplies a representation in lower dimensional spaces. In matrix terms, we perform the generalised singular value decomposition of the matrix obtained as follows:

$$\underset{(n,q)}{\mathbf{T}} \rightarrow \underset{(o,q)}{\mathbf{P}} = \underset{(o,n)}{\mathbf{Z}'}\ \underset{(n,q)}{\mathbf{T}} \rightarrow \underset{(o,q)}{\mathbf{F}} = \underset{(o,q)}{\mathbf{P}}\ {}_c\underset{(q,q)}{\mathbf{D}_p^{-1}} \rightarrow \underset{(o,w)}{\hat{\mathbf{F}}} = \underset{(o,q)}{\mathbf{F}}\ \underset{(q,w)}{\mathbf{X}'(\mathbf{XX})^{-}\mathbf{X}} \rightarrow \underset{(o,o)}{\mathbf{D}_o^{-1}}\ \underset{(o,w)}{\hat{\mathbf{F}}}$$

7. Characteristics of the use of lexical structures with variable elements during the Nineties in "La Repubblica"

The matrix on which we have based the analysis has dimensions (10,10). It has the symbolic objects previously defined in rows and the years in columns. Therefore, the generic element is given by the frequency of a complex lexical structure conditioned by the symbolic object to which it belongs (see Tab. 3, for the absolute frequencies).

Thus, the analysis performed on the profile matrix **F** is the analysis of the dependence of lexical structures on the time when they were used. It is based on the decomposition, in factorial terms, of the τ_b index proposed by Goodman and Kruskal (Lauro, D'Ambra, 1984).

Tab. 3 - The **P** *matrix: symbolic objects in the ten years considered*

Symbolic Objects / Years	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	Mean Corpus
Left wing	1115	963	812	689	474	423	444	492	513	504	639
The Right	235	263	318	343	463	322	300	294	258	266	309
The Centre-Liberal	103	91	93	126	185	155	141	142	154	157	135
Religions	37	33	37	36	43	42	37	48	43	52	41
East	155	220	148	166	140	122	164	144	173	148	157
West	70	76	67	70	60	59	76	72	74	90	71
Terrorism	189	234	188	235	205	208	205	212	220	196	209
Famous people	43	32	39	27	34	27	25	26	39	44	33
Philosophical movements	24	18	16	15	16	13	14	17	12	14	16
Separatism	51	53	53	59	92	77	115	86	59	56	71

Legenda: Normalised occurrences (per million)

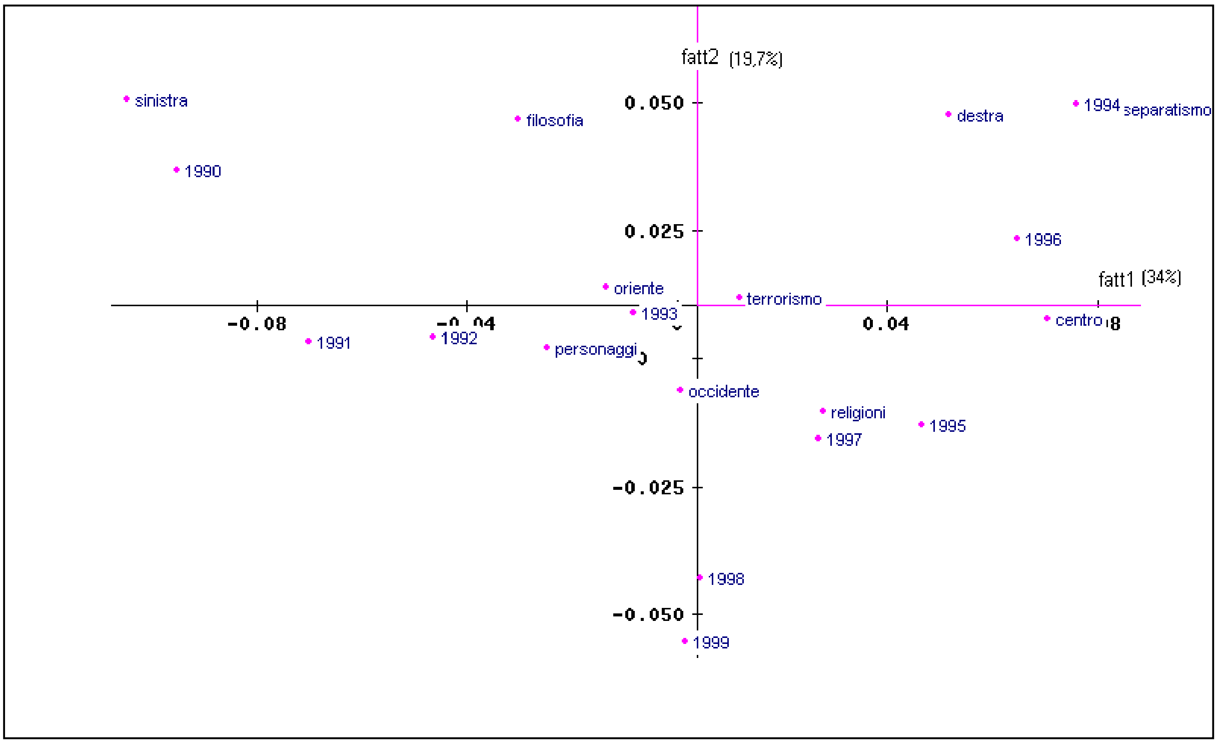
About 54% of the information on the studied dependence structure is shown on the first factorial plane and it sums up to 70%, if we consider the third axis (*see* Tab. 4).

Tab. 4 – The eigenvalues of the non symmetrical analysis

Eigenvalue	%	Cum. %
0,17	34,0	34,0
0,10	19,7	53,7
0,08	15,0	68,7
0,05	10,9	79,6
0,04	8,2	87,8
0,03	5,5	93,3
0,02	3,5	96,8
0,01	2,6	99,4

In Fig. 3 we point out, on the first axis (34%), the opposition among the objects related to the *political context*. On the negative side we have the early Nineties, characterised by the fall of the Wall and the lexical structures related to the *Left wing*. On the positive side, we have the middle years of the Nineties, in particular 1994, when in Italy the *Right* and the *Centre* wing gained power and we have topics related to *Separatism* (the *Lega* political party).

Fig. 3 – First factorial plane: active elements (symbolic objects/years)



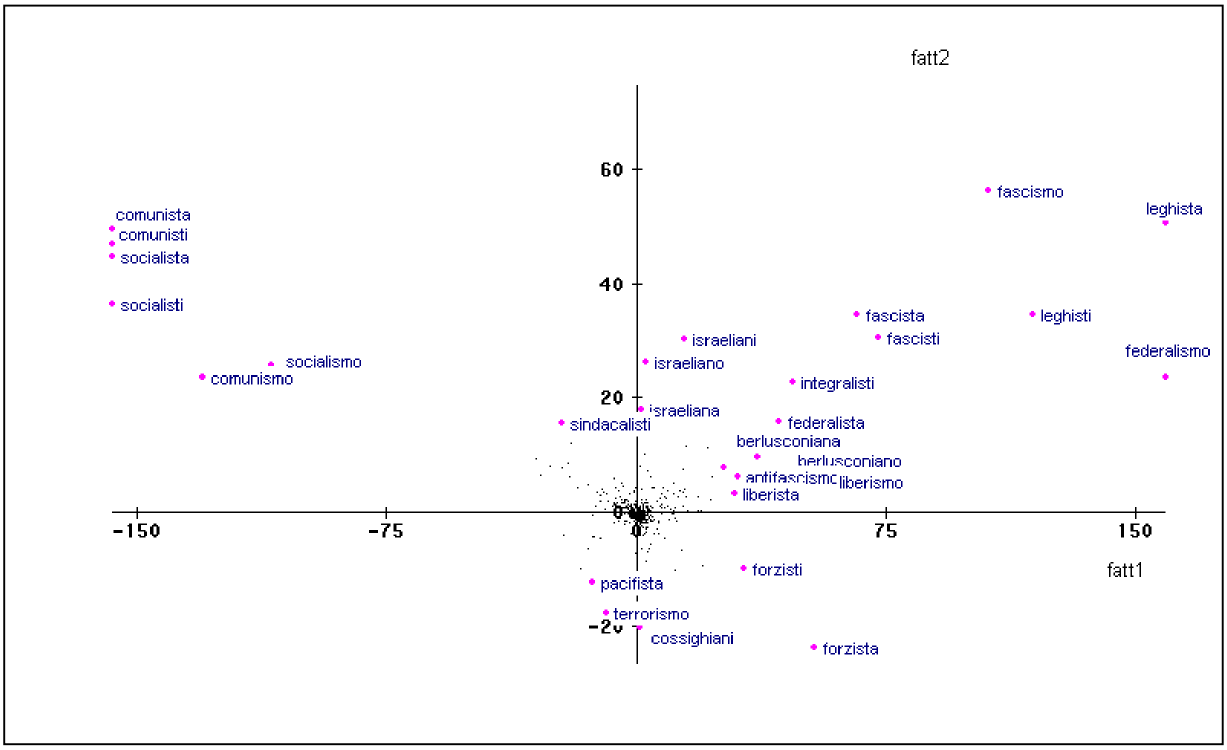
As previously said, this analysis is enriched by the lecture of the forms that have been important in this opposition. In fig. 4, we can read, on the left side, the words related to the Left wing ideology (e. g. related to the bases *social-* and *communist-*[10]) having very high

[10] Let's label *-social-*, *-comunist-*, and so on, all the elementary flexions related to those lemmas that are very near in the graph.

negative co-ordinates (points are shifted to the frame of the graphic). These words highlight the importance of topics related to international subjects, at least concerning with these lexical structures implicitly "partisan" (e.g. *anti-*, *filo-*). Therefore, on the left side we have words about the Arabian world (*-iranian-*, *-sirian-*, *antiamerican-*) probably related to the Gulf War. On the graph right side, we have words related to domestic politics, in opposition to the world events of the previous period (e.g. *liber-*, *federal-*, *leghista/i*, *forzista/i*, but also words with *fascis-* base and adjectives as *berlusconiano/a*). The second factorial axis shows, in the lower part, the most recent years characterised by topics not related to political themes (e. g. *Famous people*, *West*, *Religions*) in oppositions to the other years characterised by political and ideological subjects (*Left wing*, *the Right*, *Philosophical movements*, *Separatism*).

Jointly observing figures 3 and 4 we can clearly see the contribution that this method can supply. We point out that forms connected with terrorism have a central position (thus not very interesting because widely use in all the ten years). On the contrary, the word *terrorism* is typical of the articles published in the late Nineties.

Fig. 4 – First factorial plane: supplementary elements (words)



References

Balbi S. (1995). Confidence regions in factorial representations for textual data with non symmetrical correspondence analysis, in S. Bolasco, L. Lebart, A. Salem (eds.) *JADT'95*, III Giornate Internazionali di *Analisi statistica dei dati testuali*, Roma, tome 2, 5-12.

Balbi S. (1997). Graphical Displays in Non Symmetric Correspondence Analysis, in J. Blasius & M. Greenacre (eds.) *Visualization of Categorical Data*, Academic Press, San Diego (CA, USA), 297-309.

Balbi S., Giordano G. (2000). Un'analisi dei dati testuali con informazioni esterne: le definizioni di qualità, in M. Rajman & J. C. Chappelier (eds.) *JADT 2000 5es Journées internationales d'Analyse statistique des Données Textuelles*, Ecole Polytechnique Federal de Lausanne, Lausanne 339-345.

Biber D., Conrad S., Reppen R (1998). *Corpus Linguistics*. Cambridge University Press,Cambridge, pp. 295.

Bock, H., Diday, E. (2000). *Symbolic Data Analysis*, Springer Verlag.

Bolasco S. (2000). TALTAC: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique. in M. Rajman & J.C. Chappellier (eds.) *JADT2000*, EPFL, Mars 2000 Lausanne, tome 2, 342-353.

Bolasco S., Verde R. Balbi S. (2002). *Outils de Text Mining pour l'analyse de structures lexicales à éléments variables* (in this issue).

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus*. Armand Colin/Masson, Paris, pp. 240.

Halliday M. A. K. (1994). *Lingua parlata e lingua scritta* La Nuova Italia, Firenze.

Lauro N. C., D'Ambra L. (1984). "L'analyse non symétrique des correspondances", E. Diday et al. (eds.), *Data Analysis and Informatics*, III, Amsterdam, North-Holland, 433-446.

Lebart L., Salem A., Berry L. (1998) *Exploring Textual Data*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Voghera M. (1992). *Sintassi e intonazione nell'italiano parlato*. Il Mulino, Bologna.