On-Line Forums: a window on language evolution

Enrica Aureli[2], Fabrizio Cristofori, Domenica Fioredistella Iezzi[1]

[1] "Tor Vergata" University, Viale Columbia, 1 – Rome - ITALY email: stella.iezzi@uniroma2.it
[2] SPSA Department – "La Sapienza University", P.le Aldo Moro, 5 – Rome - ITALY

Abstract

On-line forums offer a considerable amount of information, which can be exploited in order to analyse contemporary language. The aim of this paper is to propose a statistical method for analysing how the use of some words, that reported real life events, change ever time taking the form a new meaning remote.
Four newspapers-related forums, referring to different political areas, have been monitored for five months since September 11, 2001.
A corpus of about 25,000 contributions, categorized according to two demographic variables (sex and class of age), has been examined, using multiway data analysis techniques in order to investigate the trajectories of some specific words over the period of observation. Time series analysis and cluster analysis have been used to detect new occasions of a three way matrix.

Riassunto

I forum on-line costituiscono un importante patrimonio informativo sul linguaggio contemporaneo e consentono di individuare in anticipo alcune espressioni, desunte da episodi reali, che poi di lì a poco entreranno nel vocabolario. L'obiettivo di questo lavoro è di proporre un nuovo metodo statistico per individuare parole utilizzate in riferimento a fatti di cronaca, che costituiscono le "occasioni" di una nuova matrice a tre vie di "eventi chiave". La sperimentazione è stata compiuta su 4 forum aperti da giornali nazionali italiani con differenti orientamenti politici, nei 5 mesi successivi agli attentati dell'11 settembre.
Il corpus è costituito da circa 25.000 messaggi che sono stati categorizzati in base al sesso e alla classe d'età dei rispondenti. Le specifiche traiettorie delle parole sono state individuate mediante l'utilizzo di alcuni algoritmi per matrici a tre vie. Le tecniche di analisi delle serie storiche ed di cluster analysis hanno consentito di individuare nuove occasioni per la matrice a tre vie.

Keywords: on-line forums, textual forms, three-way data analysis, time series analysis, cluster analysis

1. Introduction

On-line forums are a new huge resource to analyse contemporary language (Anderson & Kanuka, 1997; Huffaker, 2005). Generally, a corpus derived from on-line forums is structured in sub-parts, which allows comparisons between different lexical profiles. For instance, some newspaper-related forums are politically characterized; due to the use of a nickname and to some aspects of Italian language (e.g. participles), a classification of the contributions according to sex can be considered; a partition into age classes (young, adult, aged) is also possible. For what concerns the linguistic aspects, on-line forums are a source of information about the use of our contemporary language.

The aim of this paper is to propose a statistical method for analysing how the use of some words, that reported real life events, change over time taking the form of a new meaning remote.

The September 11, 2001 attacks, we used to test our method. As well known, the September 11, 2001 attacks, often referred to as 9/11, consisted of a series of coordinated suicide attacks by al-Qaeda on that date upon the United States of America. On that morning nineteen terrorists affiliated with al-Qaeda hijacked four commercial passenger jet airliners.

After the September 11 on-line forums and newsgroups related to this event were very rich of contributions over a long period of time. This allows us to introduce a chronologic dimension into the statistical text analysis. Multiple correspondence analysis between lexical profiles helps us in explaining comparisons among categories (i.e. forum, political area, sex, age) according to textual variables; but the interpretation was static. On the contrary, taking a chronologic variable into account, we can also analyse their evolution through time.

Following the theoretical pattern of three-way data analysis, STATIS method (Lebart et *al.*, 2005) has been used on textual data in order to focus on the evolution of words and topics of discussion, to measure the changes in the structures of relation between words and categorical variables, and to find out which words and variables mostly contribute to this evolution. Finally, analysing the anomalous trajectories of some words allows us finding out a change in using these forms through time, i.e. a change of their meaning in this specific context.

2. Materials and Methods

Four newspapers-related forums, referring to different political areas, have been monitored for five months since September 11, 2001.

In particular, we have been considered the following on-line forums:

1) "Corriere della Sera"[1]: 10.283 messages (9,6 MB)
2) "L'Unità"[2] – 7.550 messages (3,6 MB)
3) "La Repubblica"[3]: about 20.000 messages (only partially exploited)
4) "Virgilio"[4]: 5.334 messages (2,9 MB)

A total amount of about 25,000 messages was collected (20 MB of text files) over a period of about four months since September, 11.

The different number of messages sent to forums by newspapers, e.g. *La Repubblica* presents a major number of messages (Table.1).

Table 1 No. daily average of messages by newspaper

NEWSPAPER	MEAN
La Repubblica	375,6
L'Unità	154,0
Virgilio	101,4
Corriere della Sera	82,9

[1] The web site is www.corriere.it.

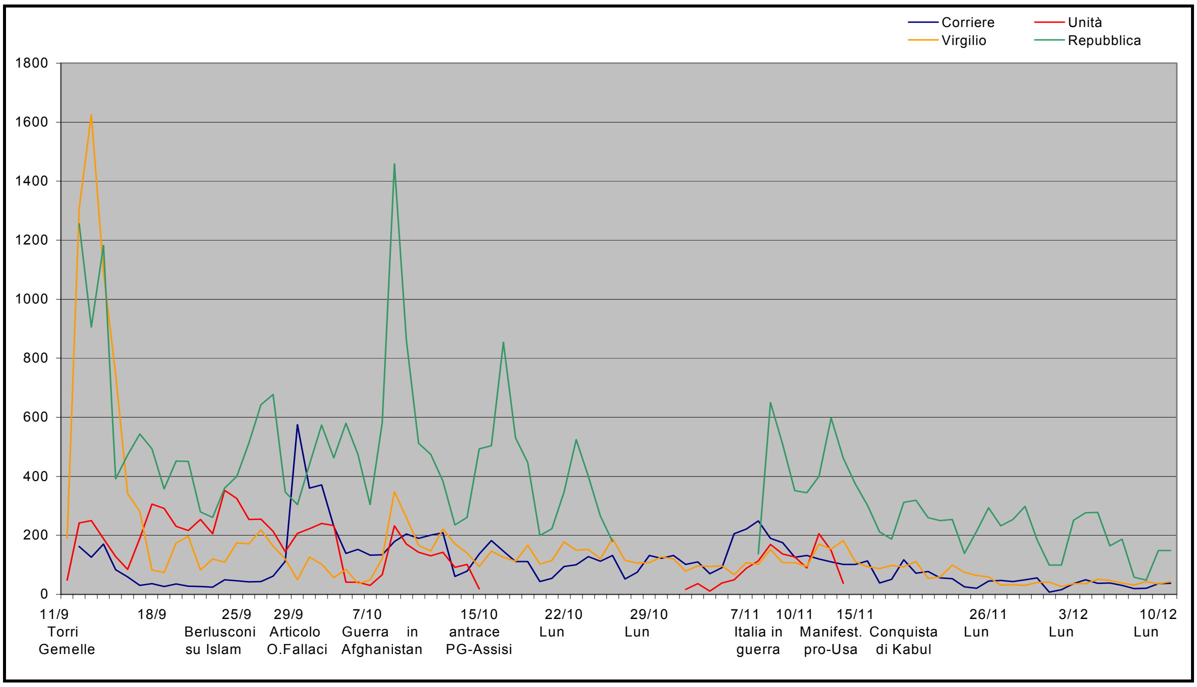
[2] The web site is www.unita.it

[3] The web site is www.repubblica.it

[4] The web site is : www.virgilio.it.

Despite a different number of messages, the four newspapers present a similar trend to close key-events.

Figure 1 *No. Messages sent to forums by date and newspaper*



We have applied a sequential method composed by the following steps:
1) Pre-processing

2) Partitioning the collected corpus into time intervals

3) Three-way data analysis.

In the first step, we cleaned and normalised the *corpus (Bolasco et al.,* 2004).

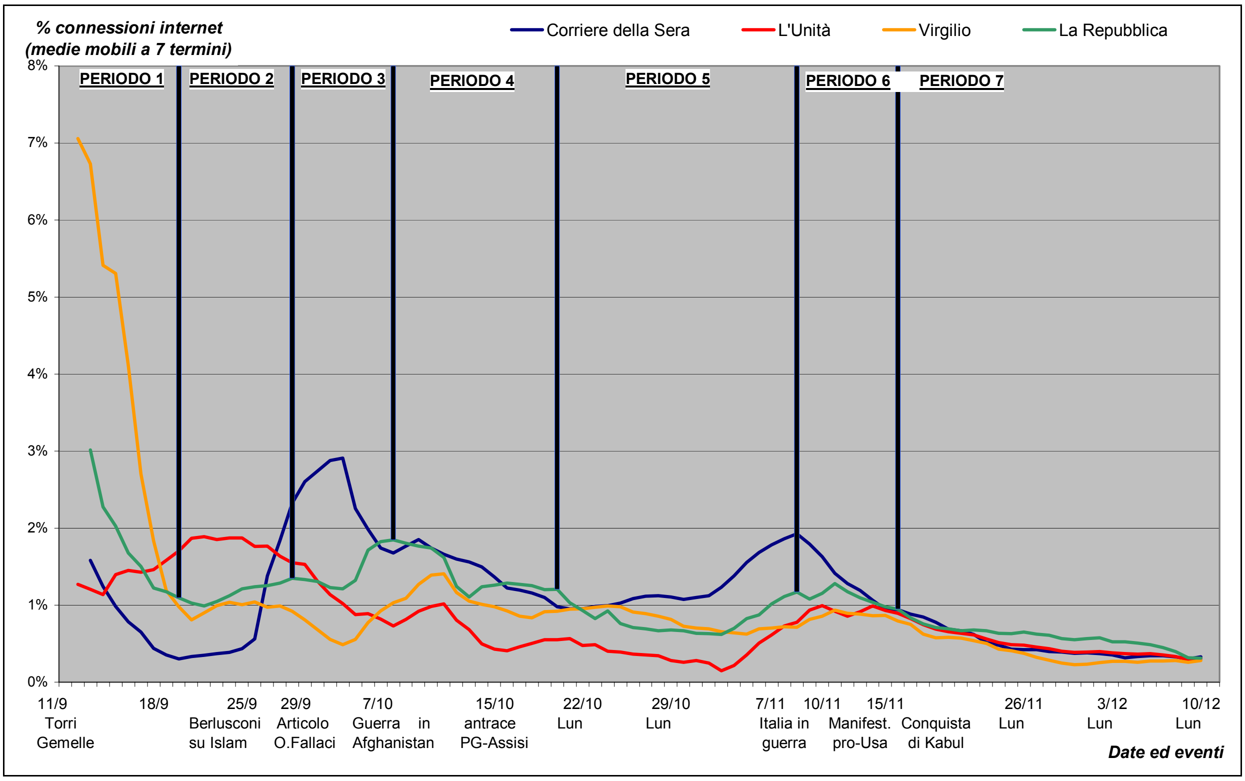
2.1. Partitioning the collected corpus into time intervals

In order to proceed with the three-way data analysis, a preliminary partition into time intervals was necessary. As far as a partition into weeks or months didn't really reflect the evolution of the discussion in the forums, two different methods have been followed and then integrated each other.

2.1.1. Partition based on the level of participation in the forums by key-events (time series analysis)

A first partition has been obtained examining the evolution of the daily number of contributions in the on-line forums. Time series analysis based on 7 terms moving averages was applied. Some anomalous dates have been detected. These are the days when the participation in the forums is significantly higher than usual. Events causing such a high participation are considered to be *key-events*, e.g. 7/10 first bomb attack on Kabul, 15/10 peace march Perugia-Assisi, 7/11 voting of Italian Parliament for taking part in the military missions (Figure 2).

Figure 2 Partition of events based on time series analysis

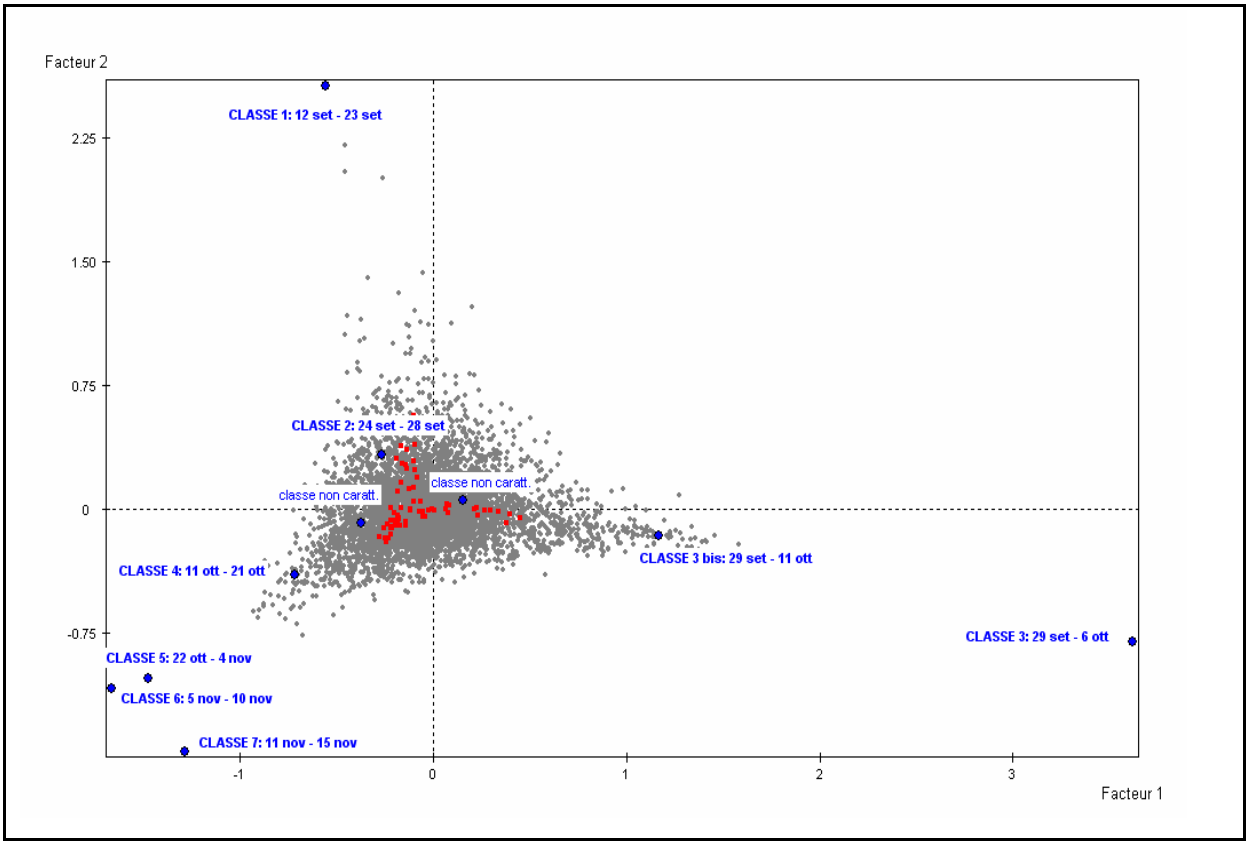


2.1.2 Partition based on a cluster analysis of textual data (mixed classification algorithm)

The so-called key-events induce not only a higher level of participation, but probably also a change of content in the contributions. That's why our purpose was to obtain a partition of the corpus in time intervals which could be coherent with the text contents of each interval. In order to achieve this purpose, a cluster analysis of textual data has been carried out. Textual data have been considered as statistic units of analysis and have been categorized according to the dates. Cluster analysis has allowed identifying groups of days (periods) which are homogeneous in relation to the content of contributions in the forums.

The procedures of class characterisation through nominal variables and the positioning of centroids on the first factorial plan have helped with the interpretation of the classes. Procedures SEMIS, PARTI and DECLA of SPAD 5.0 have been used (Figure 3).

Figure 3 Positioning of centroids on the first factorial plan – (forum Corriere della Sera)



2.2. Three-way data analysis

The *corpus* is now partitioned according either to some categorical variables (sex, age, forum, political area) or to time intervals. This means that, for each time interval, we have a sub-corpus which is partitioned by categorical variables. In other words, we can analyse a sub-corpus according to these variables; this analysis can be repeated as many times as many intervals we have.

Let $\boldsymbol{X} \equiv \{\boldsymbol{n}_{ijt} : \boldsymbol{i} \in \boldsymbol{I}, \boldsymbol{j} \in \boldsymbol{J}, \boldsymbol{t} \in \boldsymbol{T}\}$, be the three-way data matrix array, where n_{ijt} is a frequency of the j^{th} messages observed on i^{th} word according to the h^{th} situation, and $\boldsymbol{I} \equiv \{1,...,\boldsymbol{n}\}$, $\boldsymbol{J} \equiv \{1,...,\boldsymbol{k}\}$ and $\boldsymbol{J} \equiv \{1,...,\boldsymbol{k}\}$ and $\boldsymbol{T} \equiv \{1,...,\boldsymbol{T}\}$ are the sets of indices of modes *i* (word), *j* (message), and *t* (key-event), respectively.

The analysis has been carried out following the theoretical pattern of STATIS method (Lebart et *al*. 2000). Our main purpose is to focus on the common structure of the various tables, and also to study the deviations of each table from this common structure.

1) comparisons among tables (analysis of interstructure)

2) comparisons among clouds of words (analysis of interstructure and intrastructure, analysis of trajectories)

3) comparisons among clouds of categorical variables (analysis of interstructure and intrastructure)

Finally, some textual forms with anomalous trajectories have been detected through the algebraic decomposition of the distances among clouds in contributions by row elements. The algebraic decomposition allows identifying those row elements (i.e. textual forms) which evolve differently from one occasion to another. In order to know the direction of this evolution, a factorial analysis of the 2-dim table composed by juxtaposing all the $\mathbf{X}_t$ tables has been carried out. The anomalous trajectories of textual forms have been displayed on the first factorial plane.

3. Discussion and conclusions

Trajectories of some textual forms are here displayed. Some categorical variables are also displayed on the first factorial plane, which can help to illustrate the evolution of these trajectories. These categorical variables identify two different groups along the first factorial axe. On the left side we find those who are favourable to a military intervention, critics towards Islam, belonging to a conservative political area. On the right side, those who are against a military intervention, antiamericanists, belonging to a left political area.
Some comparisons can be very interesting. For instance, we can take the textual forms "11_Settembre" and "Torri_Gemelle". Their trajectories along the five occasions are very different. Although they start from a similar initial position, very close to barycentre (occasion 1), the evolution of these two textual forms follows very different ways (Figure 4).
It can be assumed that during the first period of observation (occasion 1) the textual forms "11_Settembre" and "Torri_Gemelle" are used in a similar way by those who contribute in the on-line forums, according to their categories. In the following periods, these textual forms are differently distributed among the categories. We can say that in the following periods these forms identify two different notions, whereas in the first period they are used as synonyms. A further analysis of their local contexts might be helpful to prove this assumption (Figure 5).

Figure 4 Trajectory of textual form <11_Settembre> (September, 11)

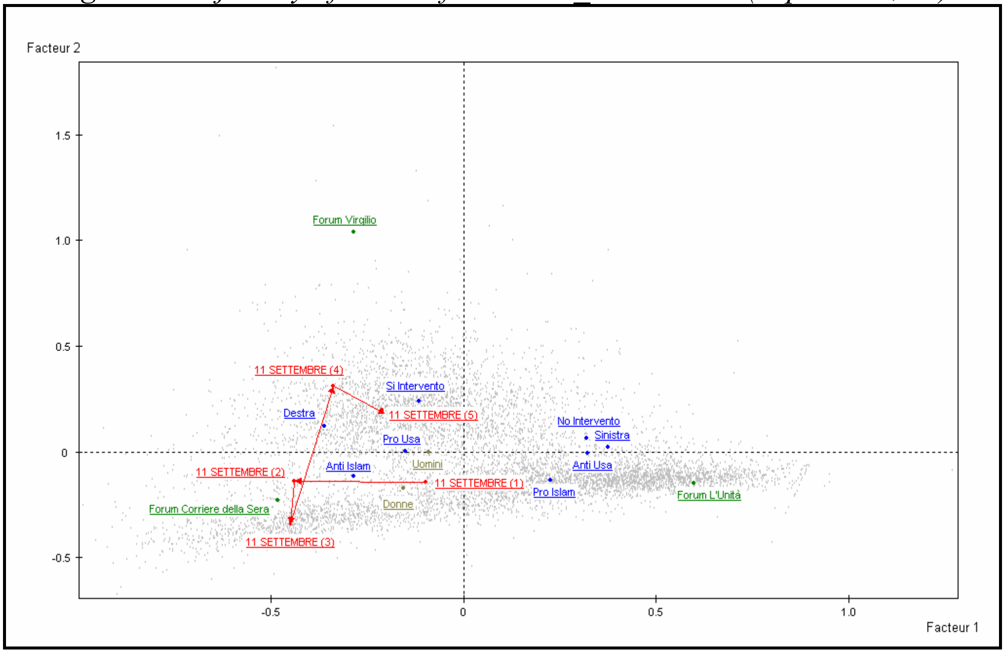
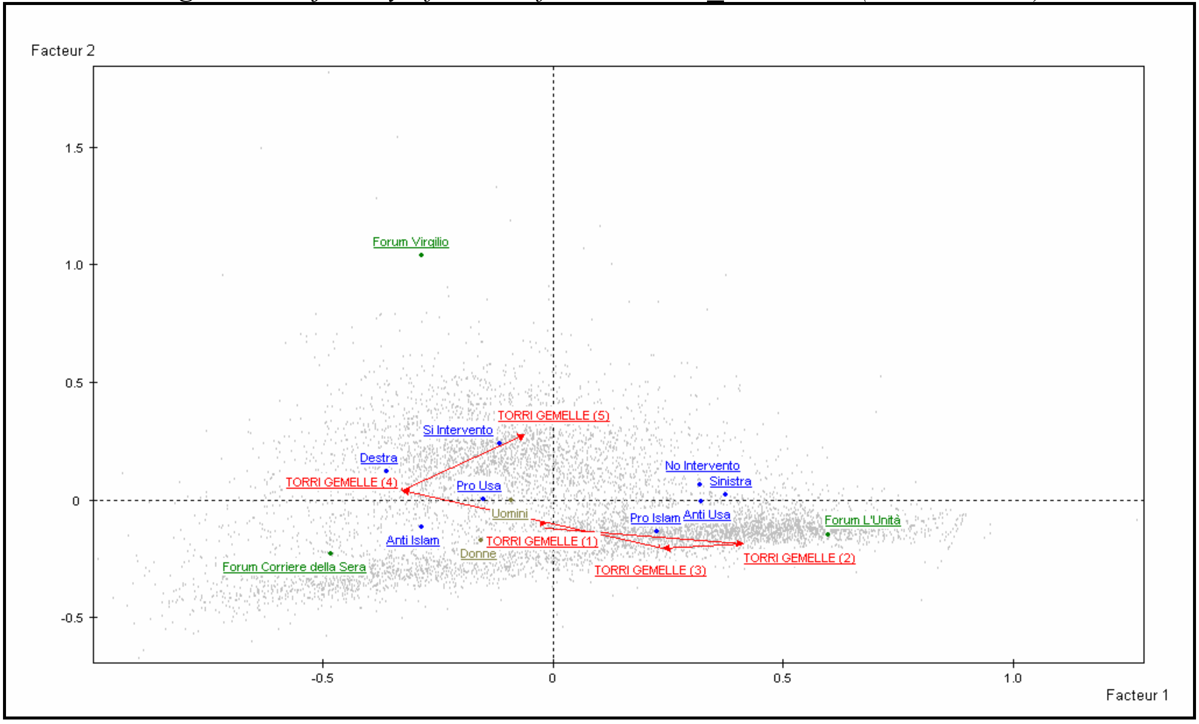


Figure 5 Trajectory of textual form <Torri_Gemelle> (Twin Towers)



The trajectories of two other textual forms, “pace” and “libertà”, are here displayed. They are not anomalous trajectories. As a matter of fact, their ways of evolution are quite similar each other. Nevertheless it can be observed that the trajectory of word “pace” is translated towards the left side of the graph, towards the “Americanisms” and those who are critical of Islamic world. On the other hand, the word “peace” seems to belong more specifically to the left and pacifist area, especially in the first two periods (Figure 6 and 7).

Figure 6 Trajectory of textual form <pace> (peace)

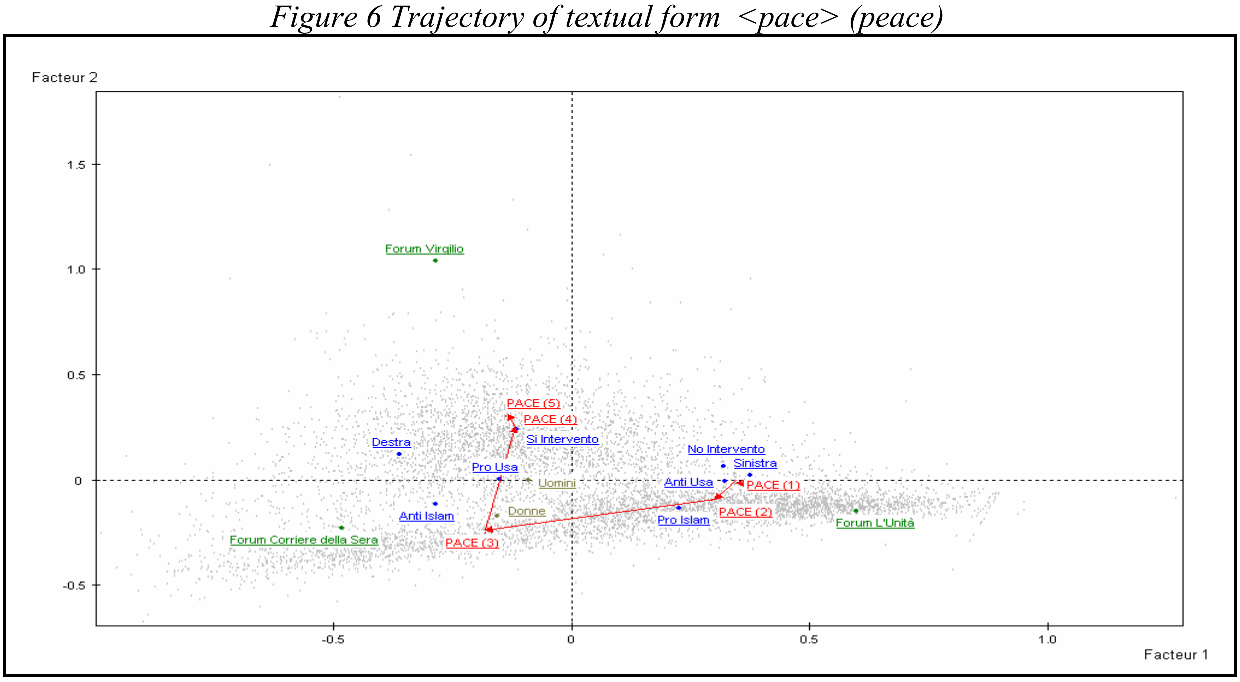
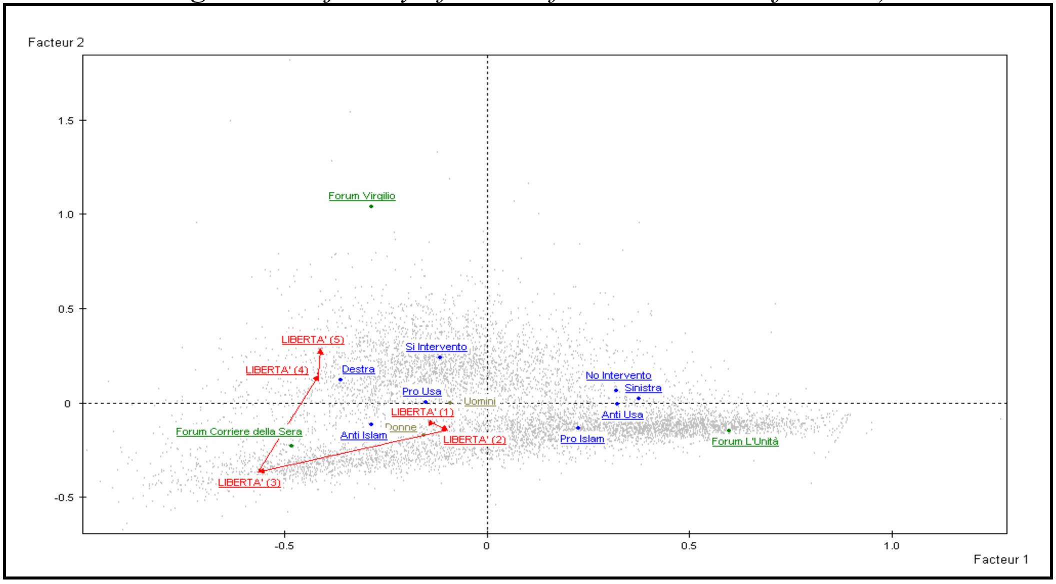


Figure 7 Trajectory of textual form <libertà> (freedom)



In this paper, a method for analysing the different use of key-words through time and detecting occasions in a three way data matrix with dynamic approach has been proposed. This method is also very useful technique to identify anomalous trajectories of words. A further analysis of the local context of these textual forms, according to categorical variables, should help to explain the different meaning involved by their different use.

5. References

Anderson, T., Kanuka, H. (1997). *Evaluating the workplace center on-line forum: Knowledge construction and learning communities.* Unpublished Research Report. Office of Learning Technologies, Human Resources, Canada.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S., Baiocchi F. e Morrone A. (2000-2003). *TALTAC 1.6. Trattamento Automatico Lessico Testuale del Contenuto*. CISU.

Bolasco S., Bisceglia B., Baiocchi F. (2004). Estrazione di informazione dai testi in *Mondo Digitale*, III, 1, 2004, p. 27-43.

Escofier, Pagès (1998). *Analyses factorielles simples et multiples.* Dunod, Paris

Giuliano L. (2002). G8-2001 : la rivolta nel monitor. Analisi testuale dei messaggi nel newsgroup <it.eventi.g8.genova> durante gli scontri di piazza. In *Actes des JADT 2002* : 301-311.

Huffaker, D. A., and Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2), article 1. http://jcmc.indiana.edu/vol10/issue2/huffaker.html

Lebart, L., Morineau, A. and Piron, M. (1995) *Statistique exploratoire multidimensionnelle*, Dunod, Paris

Mardia K.V., Kent J.T., Bibby J.M. (1979). *Multivariate Analysis,* Academic Press

Mellet S. (ed.) (1998). *JADT 1998.* Università Sophie Antipolis de Nice, Nice.

Rizzi A., Vichi M., Bock H.H. (eds.) (1998) *Advances in Data Science and Classification.* Springer Verlag, Berlin.

Rosen D., Woelfel J., Krikorian D. e Barnett G.A. (2003). Procedures for Analyses of Online Communities. *Journal of Computer Mediated Communication*, vol. (8/4). http://www.ascusc.org/jcmc/vol8/issue4/rosen.html.