Experiments on semantic categorisation of texts: analysis of positive and negative dimension

Sergio Bolasco, Francesca della Ratta-Rinaldi

Dip. Studi Geoeconomici Linguistici, Statistici, Storici per l'Analisi Regionale,
Università degli Studi di Roma "La Sapienza",
Via del Castro Laurenziano 9 – 00161 Roma – Italy
sergio.bolasco@uniroma1.it, francesca.dellaratta@uniroma1.it

Abstract

The aim of this work is the construction of a tool to categorise some properties of a text considering all evaluative words that are contained in it. For example, the tone — positive or negative — of a text can be deduced thanks to comparison with a thematic dictionary of adjectives, substantives and adverbs. The application of the dictionary to different types of languages (as dictionaries of frequency) allow us to establish a threshold point from which we can infer that a text has a negative/positive connotation.

To focus on various forms of negativity, we have carried out an analysis of all adjectives from sample of 2000 short stories by Italian students of both sexes and different age group. [1]

Keywords: text mining, categorisation, evaluation, positive and negative.

Riassunto

Scopo del lavoro è la costruzione di uno strumento di analisi che consenta di categorizzare un testo a partire dalla presenza degli elementi valutativi che vi sono contenuti. Ad esempio il tono – positivo o negativo – di un testo potrà essere dedotto grazie al confronto con un dizionario tematico di aggettivi, sostantivi e avverbi. L'applicazione del dizionario a diversi tipi di linguaggio (individuati da lessici di frequenza) consente di stabilire una soglia oltre la quale un testo può essere considerato di segno negativo. Inoltre, per focalizzare aspetti diversi della negatività si è proceduto all'analisi dell'insieme di aggettivi rilevati in un campione di 2000 di racconti prodotti da studenti italiani distinti per sesso ed età.

1. Introduction

This work is part of the field of studies tied to automatic classification of textual material by using Text Mining techniques (Sullivan, 2001). The intended experimentation will analyse the evaluative vocabulary present in some textual *corpora* to automatically characterise and classify parts of the text.

According to the usual technique of the Text Mining procedure, we will use a semantic dictionary in the Italian language as a reference model to evaluate dimension (positive *vs* negative). The comparison between the dictionary and the analysed *corpus* vocabulary will

[1] Paper realised with funds "Fondi MIUR Ateneo 2002 (C26A022374)". This work comes from a common effort; paragraphs 1 to 4 are written by Francesca della Ratta, paragraphs 5 and 6 by Sergio Bolasco and Francesca della Ratta.

allow us to identify and tag evaluative terminology contained in the *corpus*, allowing automatic classification (the positive and negative dimension) of the parts that make it up.

According to the "Pollyanna Hypothesis" formulated by Boucher and Osgood (1968), cognitive psychological studies have shown that the presence of positive terminology is much more diffused than that of a negative connotation. Text Word count finds repeatedly that positive words are used far more often than negative words, among languages and cultures as diverse as Chinese, Finnish, and Turkish (Kelly, 2003). The prevalence of positive terminology is associated with a general positive tendency, identified as a basic and universal characteristic of human nature.

Furthermore, often in the history of language, positive adjectives have had longer histories than negative ones. Also, cognitive psychological experiments have shown that it is easier to learn positive language rather than negative, both for children learning their native language and adults learning a new one (Benjafield, 1992).

2. Construction of the dictionary of evaluation adjectives

To recognise and tag evaluative adjectives in a corpus, a very comprehensive dictionary must be defined. For a reference list, we will use the English language list proposed in the dictionary of the General Inquirer (GI). It was developed in 1966 by P.J. Stone, E. Kelly and others, and recently integrated by D. Dunphy and by S. Di Cicco (Stone, 1997). It is a precious instrument for automatic classification of texts because it is based on the method of Content Analysis according to selection, and on the principle that says the frequency of certain key words is representative of the content of a text.

An instrument like the GI is very interesting for Text Mining procedures because the information contained within allows one to produce an analytical output to compile a semantic profile useful to evaluate the cognitive element and the efficacy of the text in terms of value, persuasion and emotions. The GI dictionary is composed of 13,000 lemmas of the English language, classified using different categories, that are referred both to socio-psychological theories of communication processes and to specific disciplines classifications, such as economy, law or politics. In the dictionary, the relevant role is of the "positive" and "negative" categories, that count 1,915 and 2,291 lemmas respectively; among these, adjectives alone are 590 for positive and 430 for negative.

In this study, we start with adjectives, which are considered to be the most important grammatical element to define the evaluative terminology (Marchand, 1998: 108).

To construct the dictionary, adjectives positive and negative present in GI have been translated into Italian[2] creating a list of more than 1000 lemmas. This list has been integrated with other 422 lemmas of positive and negative adjectives extracted from 6500 adjectives contained in Rep90 (corpus of 10 years of the newspaper "La Repubblica"[3]). From this list, using the dictionary present in the Taltac[4] program, all the possible adjective inflected forms

[2] The dictionary used for the translation is available in the Babylon program for translation (www.babylon.com). Throughout the translation, plurality of translation possible for each adjective has been taked into account. The translation was made at an early stage by the same person, but the list was later checked by an English mother tongue expert who is working on expanding the list of other grammar categories (nouns, verbs, adverbs).

[3] About the "La Repubblica" database see Bolasco and Canzonetti (2003).

[4] Taltac is a program for automatic lexico-textual treatment for content analysis, see Bolasco *et al*. (2000).

have been listed, which has created a dictionary of evaluative adjectives that contains about 6000 different forms.

In some cases such a list may contain some ambiguous elements whether grammatical or semantic. For instance, the word *assassino* (murderer) included in the list, may be used either as an adjective or as a noun or a verb. Similarly, the word *pure* may indicate the concept of purity or, the common conjunction synonym for also.

If the grammatical ambiguity can be overlooked, especially in view of the development of an extended version of the dictionary which will include nouns and verbs, semantic ambiguity is more complex, especially in the case of positive terminology which is often used with a neutral meaning. In the development of the dictionary it would be advisable to include terms that are potentially ambiguous, unless one can envisage how to check the list of recognised terms so as to exclude, after checking the concordance, those terms (the more frequent ones) which are incorrectly classified as positive or negative. However, in long texts, the error produced by an ambiguous classification is negligible.

3. The corpus analysed

The main corpus that was used to test the present dictionary came from the nation-wide writing competition held by the State Police in Autumn 2001 that had the theme of a story entitled "And at a certain point, the police arrived"[5] . Approximately 2,000 elementary and middle school students (between 7 and 18 years of age), from all over Italy, participated in the contest. The texts (called the Police corpus from here on) made up of all the compositions written by the children counts about one million of occurrences of words (N) and its vocabulary (V) count 47,000 words. To subject this text to the list of adjectives translated by the GI may furnish a preliminary indication of the kind of evaluative connotation in the text, and therefore it may inform us of the children's image of the police.

Besides this, it evaluates the dictionary's potential to classify texts and was be tested on different types of corpora, different writing styles, contexts and sizes to verify Boucher and Osgood's hypothesis.

4. Findings

The first step of experimentation was the comparison between the dictionary and the Police corpus' vocabulary.

The results were apparently surprising: the negative terms were more prevalent than the positive ones (with a ratio of negative to positive of 114%). This result, if the Pollyanna hypothesis is considered valid, marks a strong anomaly in the text. It is classified as having a strong negative component.

This result can probably to be attributed to noticeable structural characteristics of the text that generally start with the description of a criminal event solved by the quick intervention of the police. It can be assumed that negative adjectives are mainly associated with the events that have provoked the action of the police, whereas the positive ones have been used to describe the presence — often effective — of the police. In order to verify this hypothesis, the list of negative adjectives, taken from the text, have been analyzed.

[5] The text is being analyzed for a project Young Researchers of the University "La Sapienza" in Rome.

The most frequent negativity dimensions, or anyway those represented here above with respect to the dictionary of frequency used as reference, can be linked to the places in which the actions take place (hidden[6], dark, abandoned — *nascosto, buio, abbandonato*); to the characteristics of the "guilty person" (delinquent, murderer, dangerous, ugly, suspect, suspicious, guilty — *delinquente, assassino, pericoloso, brutto, sospetto, losco, colpevole*) and to the characteristics of the victims (poor, dead, injured, frightened, desperate, wretched, tired — *povero, morto, ferito, spaventato, disperato, misero, stanco*).

Considering the partitioned *corpus* according to categories of authors, it is possible to carry out a correspondence analysis. In our case, sex and age were chosen variables (see table 1).

The first factor is greatly determined by the comparison between the stories invented by the younger pupils (7 – 10 years old) and those invented by the older ones (14 – 18 years old). As far as the second factor is concerned, the importance of the higher age compared with the male sex of the young authors (who are also the youngest) is still decisive.

Weight		Coordinates			Absolute contributions			Squared correlations		
		1	2	3	1	2	3	1	2	3
Women	40.28	0.04	0.04	0.04	1.5	4.2	**13.7**	0.30	0.28	0.42
Men	9.72	-0.15	-0.15	-0.18	6.2	**17.4**	**56.9**	0.30	0.28	0.42
Age 7-10	13.12	-0.39	-0.07	0.08	**54.2**	5.6	14.0	**0.93**	0.03	0.04
Age 11-13	24.12	0.04	0.14	-0.06	1.0	**37.1**	13.7	0.07	**0.80**	0.14
Age 14-18	12.77	0.33	-0.18	0.03	**37.1**	**35.6**	1.7	**0.76**	0.23	0.01

Table 1. Correspondence analysis of the corpus POLIZIA

When describing the factorial planes, the most characterising adjectives in each single quadrants are examined. In the plane F1-F2, three groups of adjectives can principally be found and concentrated in the younger boys, the adolescent girls and the older boys and girls.

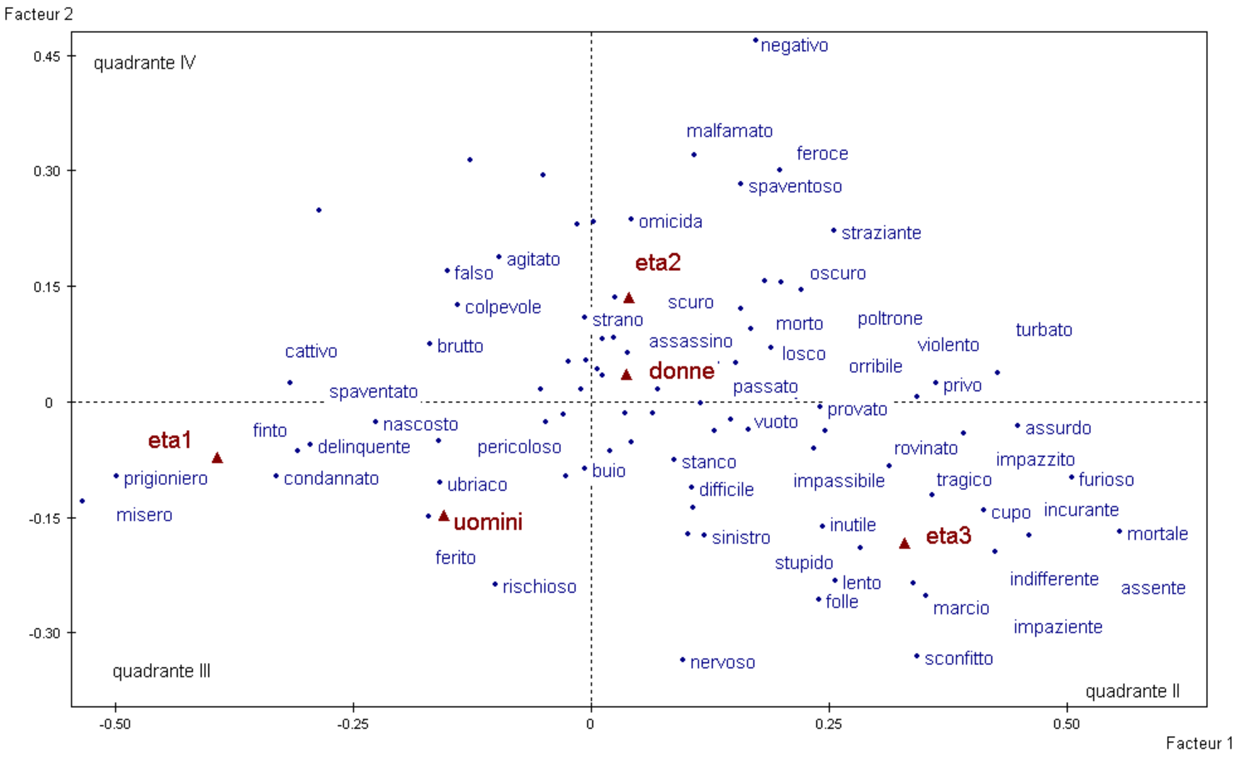
Beginning with the youngest boys, it can be noted that above all in correspondence with the quadrant III there are elementary adjectives that are used to describe the essential characteristics of the negative protagonists (bad, condemned, crafty, brusque, drunk — *cattivo, condannato, furbo, brusco, ubriaco*), the victims (prisoner, sad, injured, old, poor — *prigionero, triste, ferito, vecchio,* povero) or of the places (dark, hidden — *buio, nascosto*). These are basic adjectives, with no strong evaluative dimension, which refer to a more detached descriptive purpose from an emotional point of view. In the first quadrant, on the other hand, in correspondence with the adolescent girls, we find adjectives used for the description of the tragic situation of the victim (dead, weak, confused, upset, agonising, weary — *morto, debole, confuso, turbato, straziato, provato*) and for the elements of brutality of the negative person (terrible, murderous, assassin, frightening, evil, cruel, ferocious, negative — *terribile, omicida, assassino, spaventoso, malfamato, crudele, feroce, negativo*). In the second quadrant however, are concentrated the expressions of the older boys/girls with references to the negative protagonist, which emphasise his violent or almost demoniacal character (crazy,

[6] For clarity of description, the reference to the lemmas is given here.

mad, hateful, sinister, rotten, dirty, violent, threatening — *folle, impazzito, maledetto, sinistro, marcio, sporco, violento, minaccioso*).

The factorial analysis offers an additional key to the interpretation of the three thematic categories of adjectives (negative characters, victims, places in which the event occours), which can be linked back to the emotional dimension and varies from the description of the victim's experience to the emphasis of the "demoniacal" effects of the negative characters.

It is of particular interest to see that the youngest children are the ones who use poorer language from an emotional point of view: perhaps because the stereotypes of a fictional origin in positive and negative events are stronger in this group of children. The emotional detachment could also be a sign of the growing and often a-critical exposure to episodes of violence among the younger children, which trigger off the uncertainty of the limits between fiction and reality.



Graph 1.+ Factorial plane of negative adjectives of the Police corpus

In a similar way it is interesting to analyze the positive adjectives. These are mainly referred to **positive protagonists**, both the police forces (the policemen-*poliziotti*, but also the detective, the investigator-*investigatore*, the inspector-*ispettore*, the guard-*guardia*), and the co-protagonists, that is to say boys and girls or other characters (neighbours, teachers, passers-by, citizens — *vicini, insegnanti, passanti, cittadini*) who are given the role of protagonist in the short stories. Among the positive adjectives quoted above, those referred more often to the positive protagonists are <good, brave, strong, kind, quick in action, heroic> *(bravo, coraggioso, forte, gentile, tempestivo, eroico*). Then there are adjectives referred to **situations** which represent either the happy ending of the story or the situation at the beginning of the story that is interrupted by the criminal action (happy, tranquil, calm, quiet, carefree — *felice, tranquillo, calmo, quieto, s*pensierato) and those referred to the **objects** stolen (precious, priceless — *prezioso, inestimabile*)[2].

5. Comparison between different corpus

To fully evaluate the particularity of the results obtained, the second step of this study regards applying the dictionary to other corpora[7] types, which will show that, regardless of the type of text, positive adjectives are always more common than negative, confirming what has been affirmed by Boucher and Osgood. As it is shown in the table 2 below, the negativity index (Occ. Neg/Pos*100), in another analyzed corpora increases from 8% (in the case of the information given by the school to explain the education plan to students) to a maximum of high of 58% in the case of citizen's complaints on city services, another text with strong negative connotation. This result allows us to affirm that the dictionary is a useful instrument to characterize the negative level of a text.

CORPUS	N	V	Neg/Pos*100	Negative		Positive	
				n	**0/00 on N**	n	**0/00 on N**
1 – Police	989.785	47.354	**114,34**	13.804	13,95	12.073	12,20
2. Complaints taken from the Internet	75.361	12.587	**58,24**	866	11,49	1.487	19,73
3. Press Reviews on the World Cup 90	250.463	22.717	**52,04**	2.139	8,54	4.110	16,41
4. Focus group - temporary workers	20.691	3.460	**35,93**	83	4,01	231	11,16
5. Open questions- graduates on their thesis difficulties	48.730	4.529	**32,74**	184	3,78	562	11,53
6 Interviews with university student parents	63.548	5.545	**21,75**	244	3,84	1.122	17,66
7. Focus group teachers on nursery school	131.254	8.806	**21,44**	378	2,88	1.763	13,43
8. Open questions- graduates on satisfaction and dissatisfaction	52.684	4.702	**19,78**	180	3,42	910	17,27
9. Focus group IRRE teachers	80.558	8.154	**19,42**	209	2,59	1.076	13,36
10. Documents POF Lazio schools	92.432	10.013	**8,84**	135	1,46	1.527	16,52

Table 2. Comparison result on evaluative dictionary, ordered per negativity index

[2] The analysis of positive adjectives however presents greater problems of semantic and grammatical ambiguity, due to the natural prevalence of positive terminology which more often gives these terms a neutral meaning.

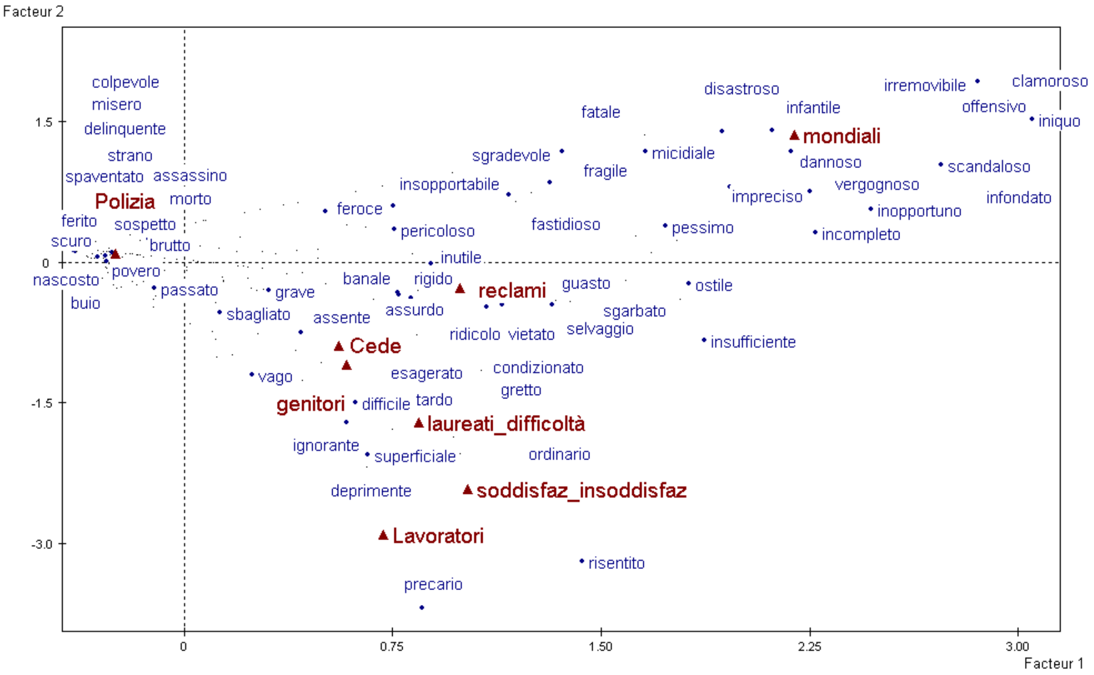
[7] The analyzed *corpora* have been collected from the writers or kindly given by other researchers. They refer to: 2) complaints taken from the Internet gathered by F. Cassoni, a university student; 3) press reviews on the World Cup '90 written by graduate students FSeM; 4) registrations of 3 focus groups on temporary workers done by Dr. L. Spera; 5) answers to open questions given to about 400 Sociology graduates in Rome in the academic years of 1997/98 on the difficulties encountered doing their thesis; 6) 30 interviews of university student parents on the future expectations for their sons; 7) 30 focus groups conducted on teachers about the quality of their nursery school; 8) the answers to open questions given to Sociology graduates on their reasons for satisfaction and dissatisfaction in their acquired training; 9) 6 focus groups on teachers on their evaluation of the nutrition education plan from the IRRE of the Lazio region; 10) documents regarding the "Plans of Educational Offer" (POF) from 16 schools in the Lazio region (10).

Furthermore, if the division of negative adjectives in the different texts undergoes a correspondence analysis, it is possible to represent the negativity dimension in some of its facets.

It is particularly interesting to note that the first factor is characterised by the opposition between the Police corpus and those of the Press Reviews on the 1990 World Cup, together with the Complaints. Such an opposition can also be brought back to the age variable, referred to the authors of the texts (the young students vs the more adult authors of the press articles). This is probably a question of the more emphasised negativity dimensions present in the texts used for the comparison, placing one in the **tragic narrative component** of the short stories on the Police (with their contents of fear, desperation, ugliness and violence) and the other in the **disgust and protest** component characterising the Press Review on the World Cup and the Complaints, with adjectives referring to scandal, negligence and rudeness, ordinariness and mediocrity, awkwardness and inappropriateness.

The second factor, on the other hand, appears to be characterised by the type of texts, with the classical comparison between written and spoken language. In fact, if the positive semi-axis is characterised by the **impersonal language** of the press, the negative one is determined by more **personal** references typical of the reviews of the focus groups, open questions or interviews. These adjectives can be interpreted as the comparison between objectivity and subjectivity: if, with regard to objectivity, there are references to harmfulness, incompetence, childishness, inaccuracy, nuisance and obscenity, with regard to subjectivity there are references to anxiety, absurdity, discontent, ordinariness, difficulty, precariousness, insecurity and embarrassment.

Of course these categories of negativity depend on the type of texts that have been analysed. In this case, it seemed interesting to demonstrate that by using the dictionary it is possible not only to order different texts according to their negativity, but also to illustrate the different negative components characterising them. If used on texts that are coherent in terms of context, such as for example political speeches or newspaper articles, this comparison could give even more interesting results with regard to the discrimination of objects.



Graph 2. Factorial plane of negative adjectives in all the corpora

After the analysis of negative dimension, the problem that has to be solved is how to define the threshold point from which on we can say a text has a negative connotation.

A possible solution could be to apply the positive-negative dictionary to some frequency dictionary in the Italian language. In this case as well, there is confirmation of the Pollyanna Hypothesis, with a negative index that varies from 50% of the enormous dictionary of “La Repubblica” newspaper to 40% of the POLIF dictionary, inserted as a reference model in the TALTAC program. This result allows us to affirm that texts with negative index higher than 40% can be considered to have a decisive negative connotation.

However index values slightly lower than 40% must be evaluated cautiously since they contain negative elements.

6. Conclusions and future perspectives

The instrument that has been defined, even though limited to the analysis of adjectives only, makes it possible to supply significant information on texts in relation to their degree of negativity. The analysis of the adjectives identified and the comparison between negative adjectives contained in different corpora then allows a more careful examination of the semantic connotation with the negative dimension. Over the coming year the dictionary will be completed with the introduction of other terms, substantives and adverbs, which will make it more complete, resolving a great part of the problems of grammatical ambiguity.

Finally, from the point of view of the automatic categorization, the possibility to tag the positive and negative adjectives in the text can permit the automatic classification of parts of the text on the basis of the greater or smaller concentration of negative terms, making it easier to extrapolate parts that present greater characteristics of negativity.

References

Benjafield J.G. (1992). *Cognition*, Prentice Hall Inc. It trad. (1995), *Psicologia dei processi cognitivi*. Il Mulino.

Bevilacqua E., della Ratta-Rinaldi F. and Orsini A. (2003). *‘Quando ad un tratto arrivò la polizia.’ Un viaggio nell’immaginario giovanile*. Progetto Giovani Ricercatori de “La Sapienza”, dipartimento DISC.

Bolasco S., Baiocchi F. and Morrone A. (2000). *Taltac. Trattamento automatico lessico-testuale per l’analisi del contenuto*. Cisu.

Bolasco S. and Canzonetti A. (2003). *Some insight on the evolution of 1990s' standard Italian, by Text Mining techniques and automatic categorization using the lexicon of the daily “La Repubblica”*. CLADAG.

Boucher T. and Osgood C.E. (1969). The Polyanna Hypotesis. *Journal of Verbal Learning and Verbal Behavior*, vol. (8): 1-8.

Hatzivassiloglou V. and McKeown K. (1993). *A Quantitative Evaluation of Linguistic Test for the Automatic Prediction of Semantic Markedness*. Columbia University. http://acl.ldc.upenn.edu/P/P95/P95-1027.pdf.

Kelly M.H. (2003). *Naming on the bright side of life*. University of Pennsylvania, http://www.sas.upenn.edu/~kellym/brightSide.html.

Kerbrat-Orecchioni C. (1981). *L’énonciation de la subjectivité dans le langage*. Armand Colin.

Krippendorff K. (1980). *Content Analysis. An Introduction to its Methodology*. Sage. It trad. (1983), *Analisi del contenuto. Introduzione metodologica*. ERI.

Marchand P. (1998). *L’Analyse du Discours Assistée par Ordinateur. Concepts, Méthodes, Outils*. Colin.

Stone P.J. (1997). Thematic text analysis: new agendas for analyzing text content. In Roberts C. (Ed.), *Text Analysis for the Social Sciences*, Lawrence Erlbaum Associates.

Sullivan D. (2001). *Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing and Sales*. Wiley.