Quaderni di Statistica
Vol. 5, 2003

Identifying word senses from synonyms: a cluster analysis approach

Carmela Cappelli
Dipartimento di Scienze Statistiche Università di Napoli Federico II
E-mail:carmela.cappelli@unina.it

Summary: In this paper we focus on synonymy relations between words. A cluster analysis approach is presented, aiming at detecting groups of synonyms of a given term which are characterised by a high degree of homogeneity and therefore are interchangeable. Some applications to the case of Italian words are shown and discussed. The results show that the proposed approach is promising in identifying different senses of a word.

Keywords: Word senses' identification, Synonyms, Cluster analysis, Semantic space.

1. Introduction

In recent years interest has grown in the application of statistical methods to various problem concerned with the analysis of textual data (for a review see Woods *et* al, 1986) where textual data can be meant as any linguistic message.

The scientific study of a language either written or oral involves different aspects and traditionally, a distinction can be made between *morph-syntax* that deals with rules of word (morphology) or sentence (syntax) construction, and it is therefore concerned with form and structure of a text, and *semantics*, which studies the meaning of words or sentences i.e., the content of a text.

Although in practice the distinction is not straight, statistics have proven to be useful for both purposes; examples are the works on sentence and word length (Yule, 1939; Piccolo, 1991; Corduas, 1995;

D'Elia 1998) on distributional laws for word frequencies (Sichel, 1975; Zipf, 1935), on authorship identification (Thisted and Efron 1987), or chronology (Yardi, 1946).

A very active field is represented by the application of multivariate data analysis tools (for a review see, Lebart *et* al, 1998). Actually, the usage of computational text analysis combined with multivariate statistical techniques allows new kinds of investigation particularly relevant in the social science when analysing survey responses, advertisements, political discourse (see for example, Balbi, 1998; Bolasco, 1996).

In this paper we focus on a strictly semantic issue: the identification of word senses by means of the analysis of its synonymy relations. In fact, the correct use of synonyms is a crucial aspect of mastering a language. Cluster analysis is proposed in order to identify group structure in the set of synonyms of a given term the use of.

By defining a measure of similarity among the synonyms that express analogous synonymy behaviour with respect to the term of interest, the set of its synonyms can be partitioned into groups, each group being characterized by a high degree of internal interchangeability. The groups therefore would represent different "senses" of the word.

The paper is organised as follows: in section 2 some considerations about the motivation underlying the study of synonyms are discussed; in section 3 the proposed approach is presented; section 4 is devoted to the applications to the case of Italian words; concluding remarks follow in section 5.

2. Motivation

A procedure to study synonymy relations among synonyms of a given word and to group them by word senses can benefit several applications:

a) *translation process* – the translation process heavily relies on the usage of dictionaries and dictionaries of synonyms nowadays available in electronic form. For a given entry, i.e. a word, the

latter return a list of words sometimes ordered according to some measure of occurrence and the choice among them relies entirely on the expertise of the human translator;

b) *information retrieval* – information retrieval systems are often based on enquiry by means of a single key-word; when the system does not recognise the query term, additional related terms are needed so that the search engine can expand the query. Automatic query expansion using thesauri represents a relevant target of research;
c) *surveys* – surveys are typically characterised by open ended questions and dealing with these requires sense disambiguation as a part of semantic processing and tagging, but, disambiguation is frequently developed on an ad-hoc manual basis.

It is our belief that in all these cases, although human interaction cannot be completely avoided a thesaurus which offers the possibility to study synonyms grouped by senses might represent a relevant help.

3. Clustering synonyms

Although they are among the simplest units into which a linguistic messages can be expressed and analysed, words have usually more than one sense in which they can used, translated etc..

The idea underlying this paper is that different meanings of a word can be inferred from its synonyms and since we are concerned with detecting group structure within them, cluster analysis appears a natural candidate approach.

Synonymy has actually been studied by Ploux (2002) with the aim to match French-English dictionaries by means of correspondence analysis whereas cluster analysis has been mainly applied to clustering texts (Willet,1988).

Since any statistical textual analysis requires qualitative information to be turned into quantitative data, the point is how to cast the problem of grouping synonyms into the framework of cluster analysis.

We may think to any word as having a *semantic space* spanned by its synonyms and the problem reduces itself to describe and partition this

space. Since we want the semantic space to be as extended as possible, we need to collect fro a given entry in a dictionary, s ay word w, all the synonyms from the available sources; the latter can be manually compiled or electronic dictionaries, web sites etc…Now, define the *extended set* $S = [s_1, s_2,\ldots, s_i,\ldots, s_k]$ of synonyms of w as the set including all the collected synonyms.

For any $s_i \in S$, let us consider the set of its synonyms. Then, we can build a $(k,k+1)$ data matrix where the rows represent the elements in S and the columns are given again by the elements in S augmented by the term w itself.

By row, each synonym s_i can be represented by a sequence of 1s and 0s according to whether each term in S and the same word w is present or absent within its own list of synonyms. In other words the synonyms are investigated conditioned to the word of interest and S is the reference set to evaluate their synonymy behaviour. The reason why the entry term w is also considered among the variables, is that synonymy is not a symmetric relation, in the sense that synonyms are not always reciprocal because some are hyponyms, i.e, subordinate words characterised by a more specific meaning whereas some are hypernyms, i.e. superordinate words that are more generic than the given word.

The data matrix has the following form:

Table 1.Structure of the data matrix

	w	s_1	…	…	s_j	…	s_k
s_1					⋮		
⋮					⋮		
s_i	…	…	…	…	δ_{ij}	…	…
⋮					⋮		
s_k					⋮		

where δ_{ij} equals 1 if the j^{th} synonym of the entry term is also a synonym of s_i and equals 0 otherwise.

A similarity measure can be then specified. Several measures have been proposed for binary data and many of them differ with respect to

the inclusion of the conjoint absence of attributes, details can be found in Gordon (1999).

In this context it is our belief that the shared absence of a variable (in the following denoted by *AA*) is indicative of similar synonymy behaviour because similarity is evaluated conditioned to a given term i.e., considering its synonyms and in this respect it is qualified by what terms mean (presence) as well as by what they do not mean (absence).

Therefore we focus on the simple matching coefficient of Sokal and Michener (1958) which is defined as the ratio of matches (conjoint presence *PP* plus conjoint absence *AA*) to the total (conjoint presence plus conjoint absence plus mismatches):

$$SM = \frac{AA + PP}{k + 1}. \tag{1}$$

In general this measure ranges from 0 to 1, being 0 if there are no matches and 1 if there are no mismatches. Given the way the matrix is defined, the smallest value is 1/*k* because any word is a synonym of itself.

To give an insight into the proposed approach, let us consider the following illustrative example. Assume that word *w* has three synonyms: s_1, s_2 and s_3, then, its extended set of synonyms is $S_w = [s_1, s_2, s_3]$. Now, for each synonym in S_w let the corresponding extended set of synonyms be:

$$S_{s_1} = [s_2, w];$$
$$S_{s_2} = [s_1, s_3];$$
$$S_{s_3} = [s_1];.$$

Table 2 reports the data matrix built according to the structure defined in Table 1, and the corresponding similarity matrix obtained considering coefficient (1).

In this illustrative example the most close synonyms of *w* are $s_2$and s_3 which might be seen as hyponyms of *w* since they do not include it in their own list of synonyms, whereas synonym s_1 does.

Table 2. The data and the similarity matrix for the illustrative example

	w	s_1	s_2	s_3
s_1	1	1	1	0
s_2	0	1	1	1
s_3	0	1	0	1

	s_1	s_2	s_3
s_1	1		
s_2	0.5	1	
s_3	0.25	0.75	1

Once the similarity matrix is defined, a clustering algorithm can be applied. Agglomerative hierarchical methods seem appropriate because by iteratively merging the most similar objects they result in a sequence of clusters that are partially nested and hierarchy can reveal hyponyms and hypernyms of the given word.

Then, the inspection of the hierarchy allows to identify main senses in which the semantic space of the given term can be partitioned (high nodes in the dendrogram), to study patterns i.e., finer and finer senses corresponding to nested clusters generated at lower levels in the hierarchy and finally to identify separate senses which are connected on top of the hierarchy.

4. Applications

In order to show how the proposed approach works, it has been applied to some terms drawn from the Italian language; the extended sets of synonyms have been defined merging the five main Italian dictionaries of synonyms[1] and an online source (http://parole.virgilio.it/parole/sinonimi_e_contrari/). It must be stressed that any thesaurus is conceptual, in the sense that it groups words or word meanings into categories that reflect the author's background and beliefs with respect to the language. As a consequence we have found that the list of synonyms provided by each one does not entirely match.

In the applications we have analysed the case of a noun, *scolaro* (7 synonyms), of a verb, *piantare* (20 synonyms) and of an adjective, *solare* (16 synonyms) ; the extended set of synonyms are:

[1] The dictionaries are: A. Gabrielli (1967, Loescher), Pittàno (1987, Zanichelli), B. M. Quartu (1994, Rizzoli), G. T. De Mauro (2002, Mondadori), P. Stoppelli (2002, Garzanti).

- *scolaro*: allievo, alunno, discente, discepolo, educando, seguace, studente;
- *piantare*: abbandonare, coltivare, cessare, conficcare, collocare, ficcare, innestare, interrare, inserire, interrompere, infilare, introdurre, lasciare, mettere, mollare, porre, seminare, sistemare, smettere, troncare,
- *solare*: luminoso, splendente, brillante, sfolgorante, scintillante, raggiante, radioso, chiaro, evidente, lampante, palese, visibile, innegabile, indiscutibile, indubitabile, lapalissiano.

A crucial point in any clustering procedure is the choice of the aggregation criterion. We have considered the single linkage that evaluates similarity among clusters as the smallest distance between any pair of objects in them.

This criterion is known to be affected by a chaining tendency in the sense that if "intermediates" are present between distinct clusters, it tends to merge them despite the fact that they are well separated.

Actually, in our context, since we are searching for finer (specific) and main (broader) senses of a word this tendency seems to be rather an advantage because senses are often either chained or nested. Therefore in the following applications we focus on this criterion.

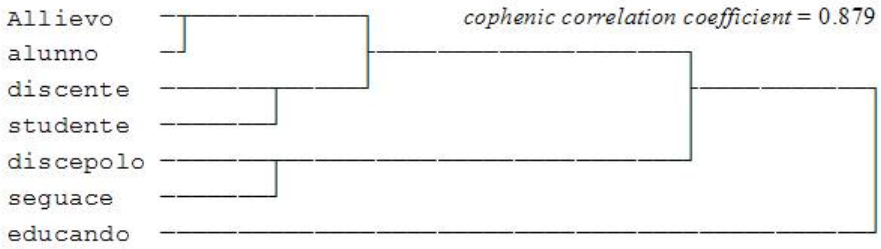

Figure 1. Dendrogram for the synonyms of noun "scolaro"

From the inspection of the dendrogram depicted in Figure 1, some remarks arise. The high levels in the hierarchy representing larger clusters identify main (broader) senses whereas bottom aggregations represent specific senses. We see that the word *scolaro* has three main senses

1) *allievo, alunno*, *discente, studente*: someone who learns knowledge and/or is enrolled in an educational programme;
2) *discepolo, seguace*: someone who takes up knowledge or beliefs from a "master";
3) *educando*: young person leaving in a college.

The first sense that can be generically meant as learner, can be further partitioned into two more specific senses: pupil (*allievo, alunno*) and student (*discente*, *studente*). Note also that the synonym *educando* appears far away from all the other, identifying always a separate sense.

The case of verbs is particularly interesting since, several senses can usually be attached to them. The dendrogram for the verb *piantare* is reported in Figure 2.

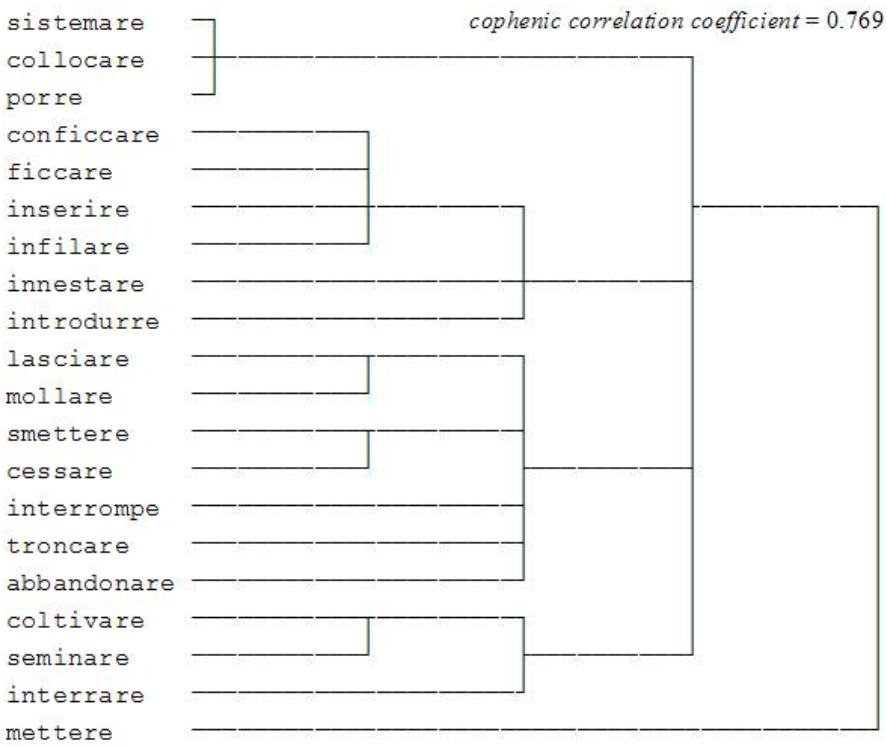

Figure 2 . Dendrogram for the synonyms of verb "piantare"

The dendrogram shows that the verb *piantare* has five main senses:

1) *sistemare, collocare, porre*: arrange or put things (into a proper order);
2) *conficcare, ficcare, inserire, infilare, innestare, introdurre:* put or introduce somethinginto something else;
3) *lasciare, mollare*, *smettere, cessare, interrompere, troncare, abbandonare*: put an end to a state or an activity, leave (someone or something), give up.
4) *coltivare, seminare, interrare*: put (seeds) into the ground;
5) *mettere*: put into a certain place.

Note how sense 5) is very generic and it is linked to all the other to a very high level; senses 2), 3) and 4) can be disaggregated into more and more finer senses whereas sense 1) is defined by very homogeneous words.

In a further application we have considered the case of the adjective *solare* (Figure 3).

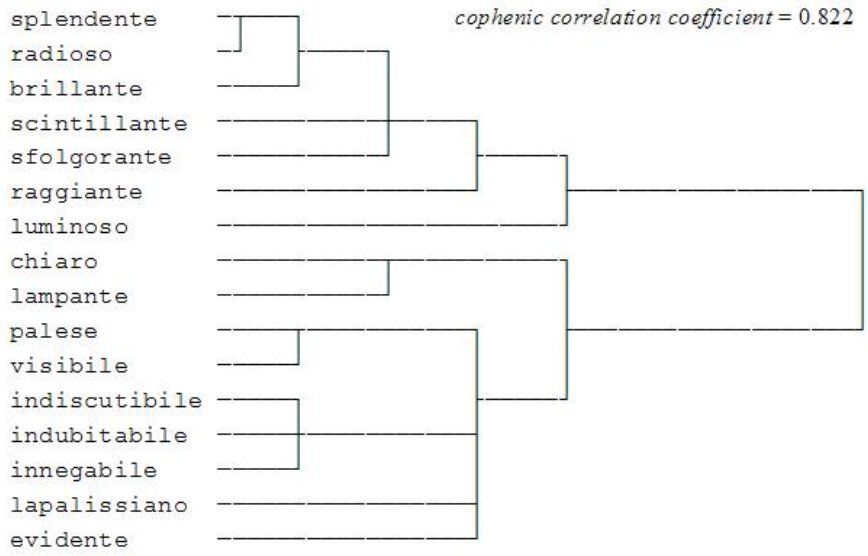

Figure 3 **.** *Dendrogram for the synonyms of adjective "solare"*

From the dendrogram two main senses stand out:

1) *splendente, radioso, brillante, scintillante, sfolgorante, raggiante, luminoso*: something abounding with (sun)light, or emitting/reflecting light readily or in large amounts;
2) *chiaro, lampante, palese, visibile, indiscutibile, indubitabile, innegabile, lapalissiano, evidente*: something clearly apparent or obvious to the mind or senses, easily perceptible, free from doubt.

For sense 1) a pattern appears that goes from *splendente* to *luminoso*, whereas, sense 2) can be disaggregated into slightly different and more specific subsenses that can be rendered as plain (*chiaro*, *lampante*) manifest (*palese*, *visibile*) impossible to question (*indiscutibile*, *indubitabile*, *innegabile)* and obvious (*lapalissiano*, *evidente)*.

For the three examples discussed, in order to evaluate the global fit of the hierarchy to the data, we have considered the cophenic correlation coefficients (Everitt and Dunn, 2001) that compare the original values in the similarity matrix with the similarities produced by the dendrogram.

In particular, this coefficient is computed between the $n(n\text{-}1)/2$ values in the lower half of the similarity matrix and the corresponding values in the so called *cophenic matrix* that is built considering, for any pair of objects, the first level at which the two object are grouped in the hierarchy. In our applications the values associated with the terms *scolaro*, *piantare* and *solare* are 0.879, 0.769 and 0.822 respectively, confirming that the data have a strong hierarchical structure.

We see that with respect to standard electronic dictionaries which provide lists based on alphabetical ordering or measure of occurrence (frequencies) ordering, the proposed procedure and its output provides additional useful information.

In particular, the output in terms of partitions offers an insight into the relations among the words in the above mentioned list to the human translators, simplifying their work and improving their productivity.

For what concerns information retrieval systems, it might be problematic for users to express their needs and translate them into queries; if the system does not recognise the query term or gives

unsatisfactory results, the aggregation programme which progressively groups its synonyms can guide the systems through the process of disambiguating and/or expanding query terms.
Finally, in surveys, traditionally open ended responses were coded manually, question by question. Nowadays software is available for cleaning and filtering in such a way as to focus on key-terms; synonyms grouped by sense might help to further select among important words to create meaningful categories.

5. Concluding remarks

We have suggested a procedure that, for a given word, turns qualitative information contained in manually compiled thesauri into a binary data matrix. This can be then analysed with statistical methods.

In our case we have considered hierarchical clustering to partition the semantic space of the word into groups. The groups present a high degree of internal homogeneity that in this context can be meant as interchangeability, each identifying a "sense" in which the given term can be meant and therefore used and translated.

The applications to some Italian words have produced promising results.

Further work will address the problem of extracting synonyms and word senses from *corpora*, i.e, collections of texts, instead of dictionaries. In fact, according to the so called *distributional hypothesis* (Harris, 1968) word with similar meanings tend to appear in similar contexts. Therefore, the study of the collocates provides the basis to define a distributionally based similarity matrix.

Acknowledgements: The work for this paper has been supported by funds granted to Dipartimento di Scienze Statistiche, Università Federico II di Napoli. The author wishes to thank M.Venuti for the helpful discussions on linguistic issues and the careful reading of the paper as well as the Academic Staff of the Faculty of Political Science for kindly participating to a preliminary test of the proposed approach.

References

Balbi, S. (1998), Lo studio dei messaggi pubblicitari con l'analisi dei dati testuali, *Quaderni del Dipartimento di Scienze Economiche e Statistiche*, 1, 155-171.

Bolasco, S. (1996), Il lessico del discorso programmatico di governo, in M. Villone, A. Zuliani (a cura di), *L'attività dei governi della Repubblica Italiana* (1947-1994), Il Mulino, Bologna, pag. 163-349.

Corduas, M. (1995), La struttura dinamica dei dati testuali, *Giornate Internazionali di Analisi Statistica dei Dati Testuali*, Roma, 345-352.

D'Elia, A. (1998), Una distribuzione per la lunghezza delle parole nella lingua italiana, *Quaderni di Statistica*, 1, 101-120.

Everitt, B. and Dunn, G. (2001), *Applied Multivariate Data Analysis,* E. Arnold, London.

Gordon, A. (1999), *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman & Hall, Boca Raton.

Harris, Z.S. (1968), *Mathematical Structures of Language*, Wiley, New York.

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

Piccolo, D. (1991), Metodi statistici per l'analisi testuale, *Quaderni di Statistica ed Econometria*, 13, 1-32.

Ploux, S. (2002) A model for matching semantic maps between languages, *Computational Linguistics*, forthcoming.

Sichel, H.S. (1975), On a distribution law for word frequencies, *Journal of the American Statistical Association*, 70, 543-547.

Sokal, R. and Michener, C. D. (1958), A statistical method for evaluating systematic relationships. *University of Kansas Science Bullettin*, 38,1409-1438.

Tyhisted, R. and Efron, B. (1987), Did Shakespeare write a newly-discovered poem?, *Biometrika*, 74, 445-455.

Yardi, M.R. (1946), A statistical approach to the problem of chronology of Shakespeare's plays, *Indian Journal of Statistics*, 7, 263-268.

Yule, G.U. (1939), On sentence length as a statistical characteristic of style in prose, with applications to two cases of disputed authorship, *Biometrika*, 30, 363-390.

Willet, P. (1988), Recent trends in hierarchic document clustering: acritical review, *Information Processing and Management*, 24, 577-597.

Woods, A., Fletcher, P. and Hughes, A. (1986), *Statistics in Language Studies*, Cambridge University Press, Cambridge.

Zipf, G.K. (1935), *The Psycologhy of language. an introduction to dynamic philology*, Houghton-Mifflin, Boston.