Quaderni di Statistica
Vol. 13, 2011

Parametric and nonparametric approaches to the semantic relationship among synonymy

Maria Iannario
Department of Theory and Methods of Human and Social Sciences
University of Naples Federico II
E-mail: maria.iannario@unina.it

Maurizio Maravalle
Department of Systems and Institutions for Economics
University of L'Aquila
E-mail: maurizio.maravalle@cc.univaq.it

Summary: In this paper we model the near-synonym lexical choice related to perceived degree of synonymy by using parametric and nonparametric approaches. Alternatively to the latent semantic analysis and to the unsupervised statistical methods for automatic choice when the context is given, the study draws from a comparative analysis of two statistical frameworks. By means of nonparametric models we synthesize the level of association in a finite number of dimensions, and we identify the problems of representing near-synonyms by developing clustered models of lexical knowledge. By parametric approach, instead, we describe the uncertainty concerning the process of lexical choices and we quantify the level of perceived semantic relationship from a set of ranked synonyms. Both of them summarize in an efficient, robust, and flexible way a semantic map of synonyms and reach similar results. These alternative approaches are introduced and discussed by means of an empirical application to an Italian verb.

Keywords: Synonymy, Principal components analysis, Clustering, Multidimensional scaling, CUB models

1. Introduction

"Lexical choice is the process of selecting content words in language generation. Consciously or not, people encounter the task of lexical choice on a daily basis - when

speaking, writing, and perhaps even in inner monologues" (Wang and Hirst, 2010). It must be informed by linguistic knowledge of how the system's input data maps onto words. This is a question of semantics, but it is also influenced by syntactic factors (such as collocation effects) and pragmatic factors (such as context).

This process of choice becomes more complicated when we consider near-synonymy. By near-synonyms we mean words that have the same meaning but differ in lexical nuances (Inkpen, 2007). Usually, they are words that are close in meaning, very similar, but not identical; not fully interchangeable, but instead varying in their shades of denotation, connotation, implicature, emphasis, or register (Di Marco *et al.*, 1993). We will not add here to the endless debate on the normative differentiation of the near-synonyms and synonyms (Cruse, 1986; Sparck Jones, 1986; Church *et al.*, 1994). It is sufficient for our purposes to simply say that we will be looking at sets of words that are intuitively very similar in meaning but cannot be inter-substituted in most contexts without changing some semantic or pragmatic aspect of the message.

In fact, each word can express several implications, connotations, and attitudes in addition to its basic 'dictionary' meaning. Word meaning is in principle infinitely variable and context sensitive; thus, a word often has near-synonyms that differ from it solely in these nuances of meaning, especially when we take it out from the context. In some cases the major differences concern "how different people use the same word" (Reiter and Sripada, 2002). It can be difficult even for native speakers of a language to carry out the differences between near-synonyms well enough to use them with invariable precision, or to articulate those differences even when they are known.

Otherwise, a word often has a small number of senses that are clearly different and probably completely unrelated to each other (homographs) but are just "accidentally" collected under the same word string or present ambiguity that the context works to remove.

As a consequence, meanings, and hence differences, can be fuzzy and this argument strongly suggests to take uncertainty into account in the selection process of words.

Moreover, choosing the wrong word can convey an unwanted implication. Thus, the risk of choosing a near-synonym that does not fit with the other words in a generated sentence (i.e., violates collocational constraints) is one of the main problem of natural language generation (NLG) system (Inkpen and Hirst, 2006) which uses symbolic knowledge of near-synonym differences and adresses the implementation of computational linguistics features. It also analyzes the word sense disambiguation (WSD), which considers the complete coverage of the range of meaning distinctions (Agirre and Edmonds, 2006), and other supervised and unsupervised statistical methods for automatic choice of near-synonyms (Inkpen, 2007).

Generally, many studies provide lexical choice evaluation, and we mention the methods based on Latent Semantic Space Models (Landauer and Dumais, 1997). Starting from the Weaver (1949) recognition who said: "statistical semantic studies should be undertaken, as a necessary primary step", many statistical frameworks estimate the degree of synonymy in texts.

Words are assumed to have a finite and discrete set of senses from a dictionary, which is lexical knowledge based. An efficient, robust, and flexible fine-grained lexical-choice process is a consequence of a clustered model of lexical knowledge. To make it work, a criterion has been formalized for lexical choice as preferences to express certain concepts with varying indirectness, to express attitudes, and to establish certain styles (Edmonds and Hirst, 2002).

A related analysis has been pursued by Cappelli and D'Elia (2004, 2011) who selected a noun and checked the similarity of resulting synonymy by clustering methods and parametric models: in their analysis, the role of dictionaries and thesauri as compared to personal evaluations is a fundamental issue.

Instead, this paper investigates different statistical strategies employed in parametric and nonparametric frameworks for analysing the semantic relationship that holds in synonymy. The task we address is the selection of the best perceived near-synonym that should be used with respect to a fairly objective benchmark (a topic word) for the empirical evaluation of a near-synonym lexical choice.

More precisely, we present different methods based on scoring the choices without context of reference. Of course, we are conscious that the context affects the meaning expressed by a word in complex ways but we use a methods that rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence (dictionary-based or knowledge-based methods). In some context, methods that eschew (almost) completely external information and work directly from raw unannotated corpora are termed unsupervised methods (Agirre and Edmonds, 2006).

Specifically, the highest scoring near-synonym will be chosen in an ordered sequence (by choosing the most frequent perceived sense). This will cause several evaluation scores which we summarize in a semantic map expressed in a single dimension (differences can be multi-dimensional). Moreover, what we consider to be the best choice is the typical usage in the corpus, but it may vary from writer to writer. To verify how difficult the task is for humans, we perform experiments with human judges on a sample of respondents. Data we will propose are based on mutual information scores concerning the degree of synonymy perceived by each candidate with respect to a reference word.

By means of nonparametric models we synthesize in a finite number of dimensions the level of association, and we identify the problems of representing near-synonyms by developing multivariate methodologies of lexical knowledge. Instead, by parametric approach we describe the probability construct leading to a rank as the result of a complex choice generated by latent components we model and summarize by a finite number of parameters.

The paper is organized as follows: after the presentation and some exploratory analysis of the experimental data set (section 2), we implement both nonparametric (section 3) and parametric approaches (section 4). A final section summarizes and discusses the main results with a particular emphasis on both approaches in order to enhance relative merits and pitfalls.

2. The experimental data set

We have collected data from a large number of people of different gender, age, education, residence, socio-economic status in order to evaluate their semantic perception of closeness of synonyms with respect to a target word. A list of synonyms of a selected word has been chosen after consulting the most recognized Italian books of synonymy and submitted to respondents in alphabetic order. So, we asked to Italian native users/speakers to rank a list of synonyms of a benchmark word on the basis of their perceived synonymy.

In related works on these topics, different words have been experienced and analyzed (Cappelli, 2003; Cappelli and D'Elia, 2004; 2006). The word chosen in this paper is the Italian verb *Piantare*, which presents high polysemy. Table 1 lists the submitted 20 Italian synonyms together with English translation/meaning.

Table 1. Synonyms of of the verb "piantare" with corresponding English translation

Synonyms	English translations and interpretations
abbandonare	to give up, to abandon, to leave, to quit, to drop
cessare	to cease, to stop
collocare	to place, to position, to arrange, to lay, to locate
coltivare	to crop, to farm, to cultivate, to till, to grow
conficcare	to knock, to stick into, to tap in, to press into, to pile
ficcare	to stick, to stuff, to put, to shove
infilare	to insert, to plunge
innestare	to bud, to engraft, to graft
inserire	to insert, to put into
interrare	to bury, to earth up, to plant, to sow
interrompere	to interrupt, to shut down, to call off, to halt, to cut off, to black out, to severe
introdurre	to insert
lasciare	to let go of, to drop, to leave
mettere	to put, to place, to set
mollare	to release, to let go, to drop
porre	to put, to lay down, to set, to put down, to place
seminare	to seed, to sow, to plant
sistemare	to arrange, to put, to place
smettere	to stop, to quit, to give up, to cease
troncare	to cut off, to break off, to severe, to interrupt

Respondents were asked to assign rank from 1 to 20 without reference to a specific context, exclusively on the basis of the strength (=*closeness*) of the perceived synonymy; no ties were allowed. In addition, for each respondent, we have collected covariates as gender, age, education, frequency of reading, use of Internet, frequency of travelling, other languages spoken, and so on. After a preliminary validation check, the final sample consists of 651 respondents with 30 information for each of them.

To give a synthetic picture of our sample data, socio-demographic characteristics

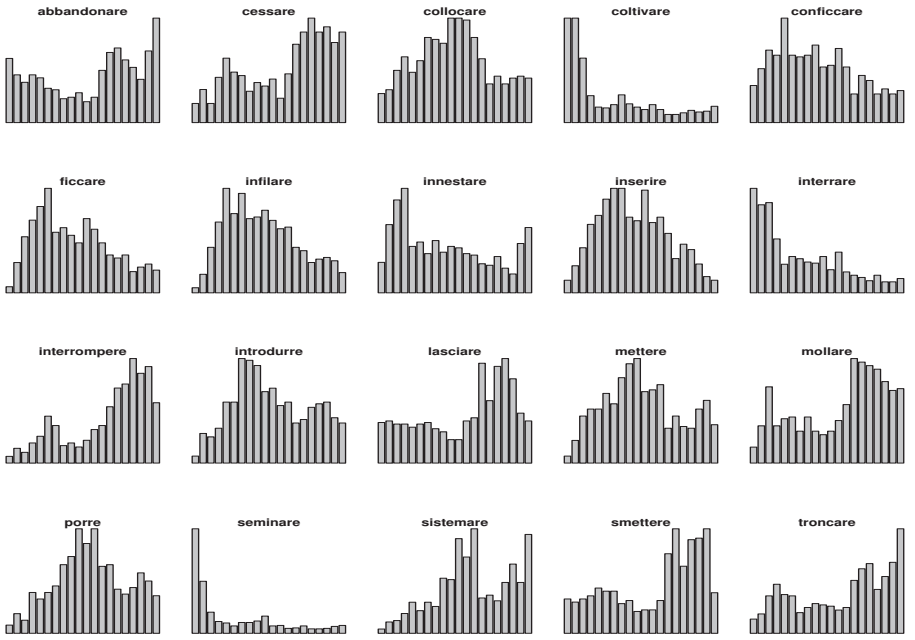

Figure 1. Frequency distributions of the ranks for the 20 *synonyms.*

may be briefly summarized as follows: 51.9% are women and 48.1% are men; the average age is 30 year but the median is 23 year so most of the population is made by young people: extreme values are two respondents of 14 and 75 years, respectively. Our sample consiste of people attending cinema 1.66 times for week on average, 60% of them connect on Internet at least once a week (survey has been planned in 2003) and read at least a newspaper during a week, the average number of cars in the family is 1.89. The sample consists of 28% of person with education up to first level, 49% received a diploma and 23% an academic degree; thus, generally speaking, the level of education in our sample is higher than the average of the population.

Collected questionnaires required a preliminary exploratory analysis and some limited imputations of missing values and/or few local corrections. At the end, the ranks of 651 respondents have been validated and they represent a complete set of subjective permutations of the first 20 integers according to the closeness of each listed synonym with respect to the target word. A complete view of the 20 frequency distributions of the expressed ranks is presented in Figure 1.

It is evident that several and contrasting shapes are present and also that some synonyms exhibit multimodality. Thus, it would be difficult to summarized all the aspects

of the distributions just with few exploratory measures. In this line, we compute a few indices of location and heterogeneity of the observed ranks and we list them in Table 2 according to some indication of *closeness* of the verbs.

Table 2. Main indicators of location and heterogeneity of synonyms for "piantare"

Synonyms	First Mode	Average rank	Gini index	Laakso-Taagepera
abbandonare	20	11.667	0.990	0.837
cessare	16	12.688	0.988	0.805
collocare	11	10.395	0.992	0.868
coltivare	2	6.243	0.939	0.436
conficcare	5	9.221	0.993	0.879
ficcare	6	9.392	0.988	0.799
infilare	5	9.711	0.987	0.794
innestare	4	9.295	0.990	0.825
inserire	7	9.625	0.986	0.784
interrare	1	6.871	0.970	0.618
interrompere	17	13.823	0.975	0.661
introdurre	7	10.828	0.990	0.832
lasciare	17	12.014	0.988	0.802
mettere	10	10.686	0.990	0.839
mollare	4	12.327	0.989	0.812
porre	10	11.525	0.987	0.792
seminare	1	5.651	0.893	0.295
sistemare	13	12.892	0.979	0.702
smettere	15	12.499	0.984	0.757
troncare	20	12.647	0.986	0.774

It is interesting to notice the position of *seminare* (the first selection for a large number of respondents) and of *abbandonare* and *troncare* (the last for many others). In addition a rough inspection of the average rank displays the semantic relationship among three clusters (approximately located near the positions 5, 9, 12, respectively). Given the ordinal nature of the rank variable we prefer to list the modal value together with the average rank whereas the heterogeneity measures are the normalized Gini $(\mathcal{G})$ and Laakso-Taagepera $(\mathcal{A})$ indices defined respectively as:

$$\mathcal{G} = \frac{m}{m-1}\left(1 - \sum_{r=1}^{m} f_r^2\right); \qquad \mathcal{A} = \frac{\mathcal{G}}{m - \mathcal{G}\,(m-1)},$$

where f_r, $r = 1, 2, \ldots, m$ are the relative frequencies of the modalities $r = 1, 2, \ldots, m$. The second index is monotonically related to the first one but its range is larger and thus it improves the discrimination among variables. In fact, for our data set, the ranges of Gini and Laakso-Taagepera indices are 0.054 and 0.463, respectively

We found that that such indices are uniformly high; in addition, the first three preferred synonyms have the minimum heterogeneity index as a consequence of a common semantic sense among respondents for these three verbs: *seminare, interrare, coltivare*.

3. Nonparametric approach

The nonparametric approach is based on the absence of a probability hypothesis about the data. In order to put this position into perspective, two of their characteristics deserve mentioning.

First, the philosophy of data analysis is founded on inductive reasoning, proceeding from particular to the general. The data set at hand and how one describes it are of importance, not the general framework or model that one might think the data fit. This standpoint has been very well summarized by Benzécri in the second principle of data analysis: *le modéle doit suivre le données, non l'inverse* (Benzécri, 1973). While few statisticians would adhere to such an extreme viewpoint, we would acknowledge that there are occasions where blind assumptions of models lead to serious defects in statistical analysis.

Secondly, from the beginning, the descriptive techniques developed by these researchers were geometric ones. Data were described to set points in multidimensional space, and points were grouped visually in a graphical display. From this point of view the goal of data analysis methods is to represent data having in mind the objective of getting an optimal synthetic representation (the so-called conceptual maps) of the information contained in the initial matrix.

In this line of reasoning, we will be concerned with three different techniques: Principal Component Analysis (PCA) and factorial methods, Hierarchical Clustering and MultiDimensional Scaling (MDS). All these analyses were made by using software `R` (2011). For other multivariate methods as correspondence analysis specifically applied to ordinal data, see Beh (2008) and Camminatiello and D'Ambra (2010), among others. Nonlinear Principal Component Analysis to discover the structure of ordinal data has been pursued by Manisera *et al.* (2010).

3.1. Principal component analysis

Principal component analysis (PCA) is a well known methodology and is one of the most important in the field of factorial methods for data analysis. The simplest interpretation is to consider it as a method to find projections of maximal variability. That is, it seeks linear combinations of the columns of the data matrix (in our case, the 651×20 matrix) with maximal variance. The first k principal components span a subspace containing the best k-dimensional view of the data. This is the classical approach initiated by Hotelling (1993) and followed by multivariate analysis textbooks

as Anderson (1958) and Mardia *et al.* (1979). Indeed, PCA may be also considered as a specific case of factorial analysis (Horst, 1965; Harman, 1967). More recently, PCA has been interpreted as a special technique for representing data which are optimal from a geometrical point of view without any reference to statistical models or probability structures, as emphasized by Lebart *et al.* (1998; 2006). This point of view may be credited to Pearson (1901) at least for the essential nonparametric content.

Table 3. Principal Components Analysis of synonyms for "piantare"

Components	Eigenvalues	% of variance	Cumulative % of variance
1	286.113705	43.949878	43.94988
2	86.100381	13.225865	57.17574
3	38.844202	5.966851	63.14259
4	28.872915	4.435164	67.57776
5	25.805410	3.963965	71.54172
6	22.384795	3.438525	74.98025
7	19.336202	2.970231	77.95048
8	17.307078	2.658537	80.60902
9	16.629334	2.554429	83.16344
10	15.625438	2.400221	85.56367
11	14.829564	2.277967	87.84163
12	13.135975	2.017815	89.85945
13	12.707848	1.952050	91.81150
14	11.891442	1.826642	93.63814
15	9.846041	1.512449	95.15059
16	8.891742	1.365859	96.51645
17	8.306067	1.275894	97.79234
18	7.512821	1.154043	98.94638
19	6.859041	1.053616	100.00000

In the case of ordinal data, several proposals of factorial methods have been introduced on the basis of multinomial distributions as discussed by Bartholomew (1980) and Jöreskog and Moustaki (2001).

In our context, we use the matrix of Spearman rank correlations. Applying this methodology to the synonymy data set, we found that the first two principal components explain 57% of the total variance, as detailed in Table 3.

Then, by using the library `FactoMineR` of `R` on the transpose of the data matrix, we obtain a display of eigenvalues (Figure 2) which permits to evaluate a synthetic performance of PCA. In fact, the first two eigenvalues are the relevant ones and thus significant interpretations may be deduced form the first two components, as depicted in Figure 3. The first component suggests a contrast between verbs denoting the idea of *breaking* (on the right) whereas the second components denotes a continuum from

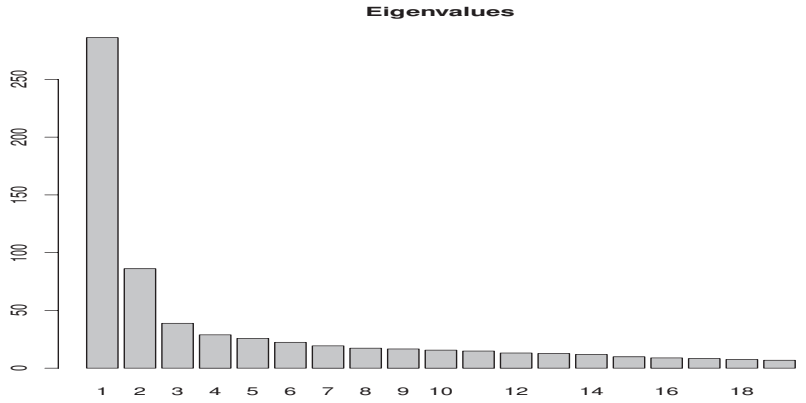

Figure 2. Distribution of eigenvalues.

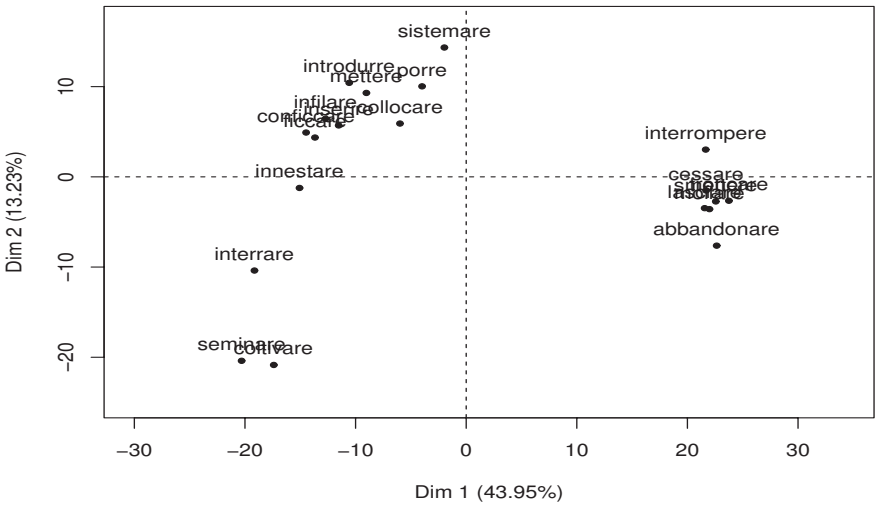

Figure 3. Verbs on first two principal components.

agricultural to situational meanings. Notice the compactness of the verbs on the left side of the representation in Figures 3.

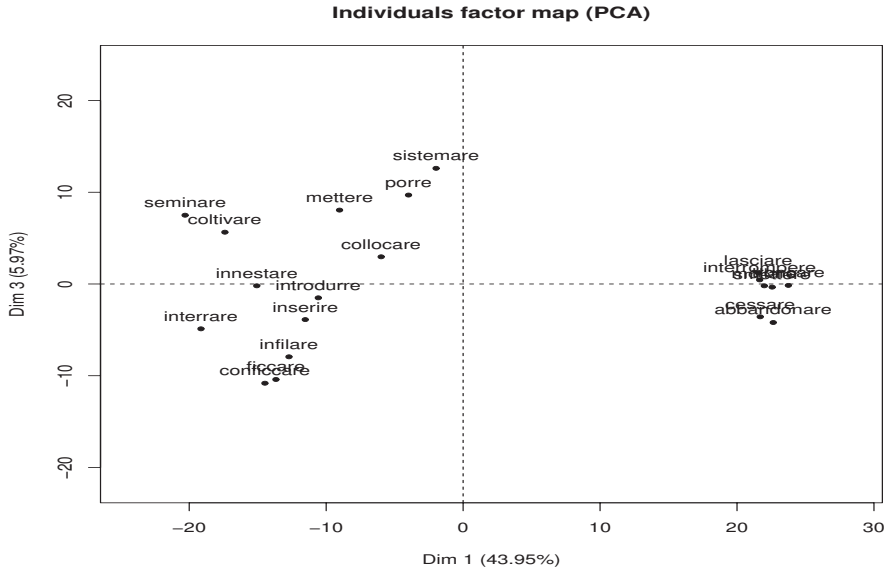

Figure 4. Verbs on components 1 and 3.

Comments that may be derived from the projections on axes (1,3) and (2,3) are not so relevant since the contribution of the third component is about 6%. Specifically, this component considers the verbs related to the idea *to insert in a hole* as opposite to the other meanings (Figures 5). Finally, the insertion of covariates in this representation (as in Figure 6) may add further information; specifically, Age seems a variables completely separate with respect to the others.

3.2. Clustering analysis

Cluster analysis is concerned with discovering groups in the original data set, and may be classified according to several criteria (agglomerative, divisive, etc.). In any case, the starting point is a similarity or dissimilarity measure by distinguishing also metric and ultrametric (Kaufman and Rousseeuw, 1990). Ultrametric dissimilarities have the appealing property that they can be represented by a dendrogram as the one shown in Figure 10; here, the dissimilarities between two different synonymy can be read from the height at which they join a single group. Jardine and Sibson (1971) argue that one method, single-link clustering, uniquely has all desirable properties of a clustering procedure.

In most of published case studies in the literature, clustering techniques assume a

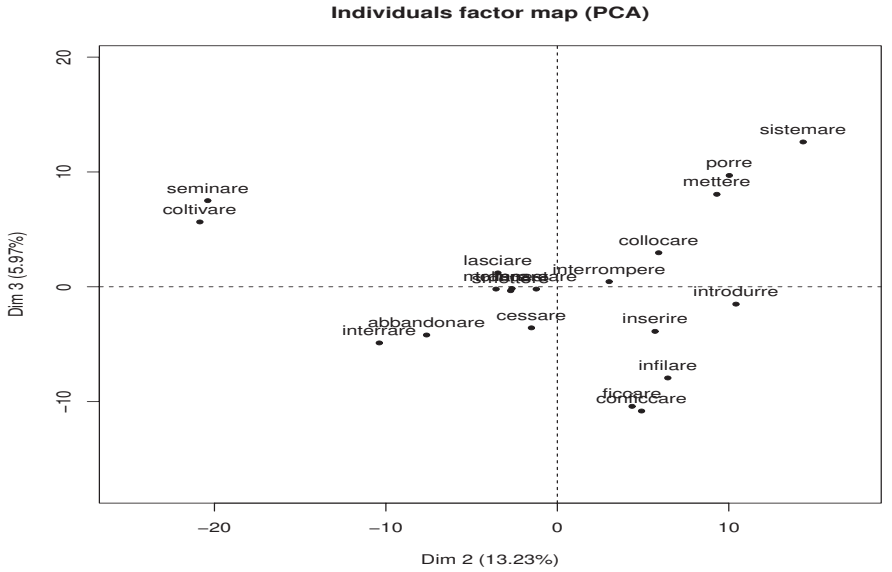

Figure 5. Verbs on components 2 and 3.

continuous range of the relevant variables; thus, their use with ordinal data is just an extension and, generally, current methodologies do not exploit the ordered structure of our data. Few proposals explicitly take the peculiar characteristics of ordinal data into account, as *Žiberna et al.* (2004), Jokinen *et al.* (2008), Giordan and Diana (2011), for instance. However, in the following, mainly motivated by computational problems, we adhere to more classical approach by using the program `pam` (developed by Rousseew) in the library `cluster` of `R`, which we apply after a convenient standardization. In addition, this choice simplifies the problems of the *a priori* determination of the number of groups since the proposed algorithm solves this problem and also the main classification objective. Thus, according to a pure nonparametric approach, data are the unique information to perform a statistical analysis.

A clustering approach to our synonymy data set has been fist pursued by Cappelli (2003). If we apply this method to the synonyms of the verb "piantare", we obtain the results represented in Figure 7 where the optimum number of groups is determined as 3.

Table 4 reports the relevant information useful to determine the three clusters as obtained by the algorithm `pam`.

It is noticeable to observe the effect of the silhouette representation obtained for the three groups as in Figure 8. Moreover, the efficacy of the cluster technique is confirmed by the substantial reduction we obtain in the sum of the within variances when we refer to 3 subgroups.

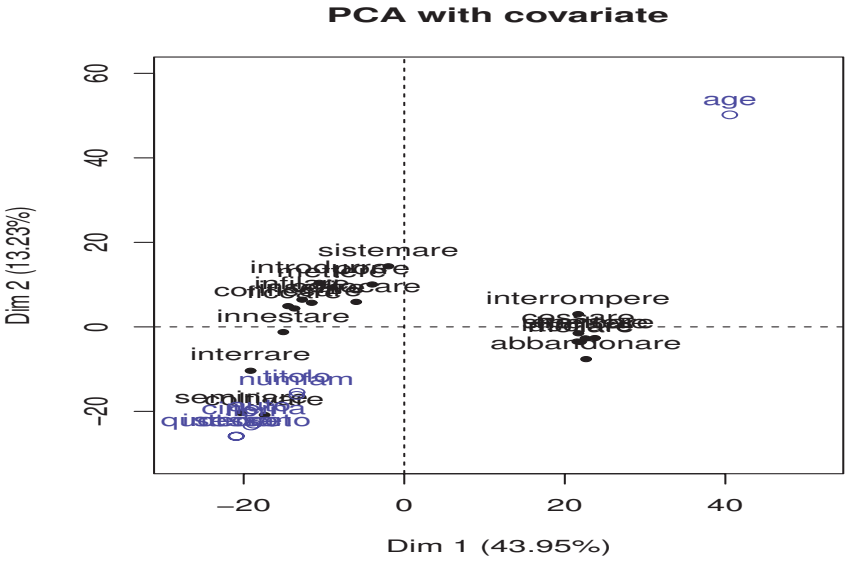

Figure 6. Verbs on first two principal components, and covariates as supplementary.

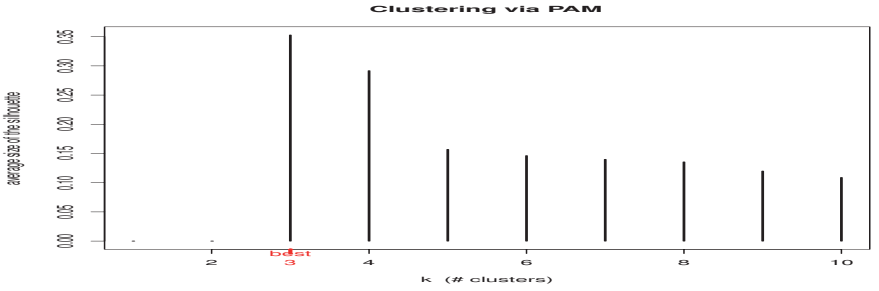

Figure 7. Optimal number of groups via PAM.

Finally, if we consider also the more classic single-linkage approach we obtain the results depicted in Figures 9 and 10 which confirm the consistency of the groups we had previously obtained.

If we jointly integrate PCA and cluster analysis we get results of even more interest. We again refer to the library `FactoMineR` of `R` that allows for a sharp visualization of both principal components and clustered subgroups (Maravalle, 2007). With regard to our data set, such representations are offered in Figures 11 and 12.

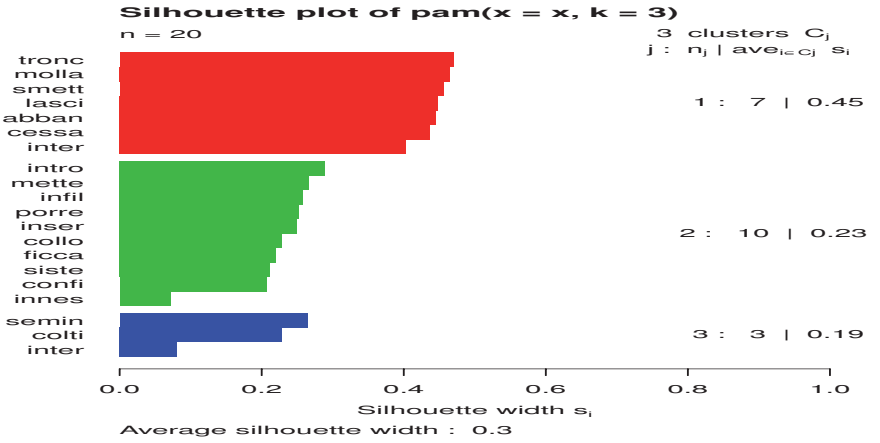

Figure 8. Silhouette for three clusters via PAM.

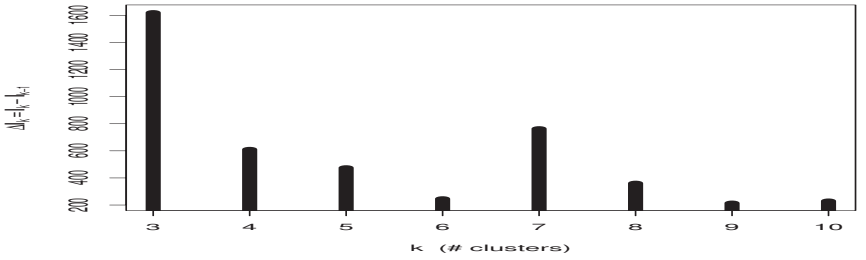

Figure 9. Optimal number of groups by k-means

3.3. Multidimensional scaling

Multidimensional scaling (MDS) techniques deal with the following problem: for a set of observed distances between every pairs of N items, find a representation of the items in few dimensions such that the inter item proximities "nearly match" the original distances. Scaling techniques were first promoted by Schoenberg (1935) and Young and Houselder (1938). Diffusion of MDS began with Torgerson (1958) and amplified with the main ideas of Gower (1966); a comprehensive catalogue is in Cox and Cox (2001).

It is impossible to match exactly the ordering of the original distances. Consequently, scaling techniques attempt to find configurations in $q \leq N-1$ dimensions such that the match is as close as possible. The numerical measure of closeness is called stress. It is possible to arrange the N items in a low-dimensional coordinate system using only the

Table 4. Silhouette plot information

	Cluster	*Neighbor*	*Silhouette width*
troncare	1	2	0.46980219
mollare	1	2	0.46525148
smettere	1	2	0.45602312
lasciare	1	2	0.44817088
abbandonare	1	2	0.44553751
cessare	1	2	0.43646983
interrompere	1	2	0.40282083
introdurre	2	3	0.28938789
mettere	2	3	0.26659051
infilare	2	3	0.25846927
porre	2	1	0.25249893
inserire	2	3	0.24931823
collocare	2	3	0.22846151
ficcare	2	3	0.22014574
sistemare	2	1	0.21148965
conficcare	2	3	0.20675662
innestare	2	3	0.07142905
seminare	3	2	0.26466762
coltivare	3	2	0.22861920
interrare	3	2	0.08092198

rank orders of the $N(N-1)/2$ original distances, and not their magnitudes. If the actual magnitudes of the original distances are used to obtain a geometric representation in q dimensions, the process is called metric MDS. Instead, when only the ordinal information is used to obtain a geometric representation, the process is called non-metric MDS, as first discussed by Shepard (1962a, b) and Kruskal (1964).

In our case study, data were first transformed in a dissimilarities matrix and then a non-metric MDS has been applied by using the program `isoMDS` of the library `MASS` in `R`. The most important results are the eigenvalues whose percentages are the following: $\{51.49,\ 14.34,\ 6.41,\ 4.40,$
$3.72,\ 3.27\}$. As a consequence, two dimensions are able to visualize the best of information contained in the distance matrix of data and this representation is shown in Figure 13. The stress is 0.070 and it can be judged very good according to the common standards of MDS analysis.

If we discriminate the results by using the covariate Age of the respondents (young and elderly are represented on the map by different fonts), we get the MDS representation depicted in Figure 14 and, for this data set, we found no relevant difference of results with respect to Age.

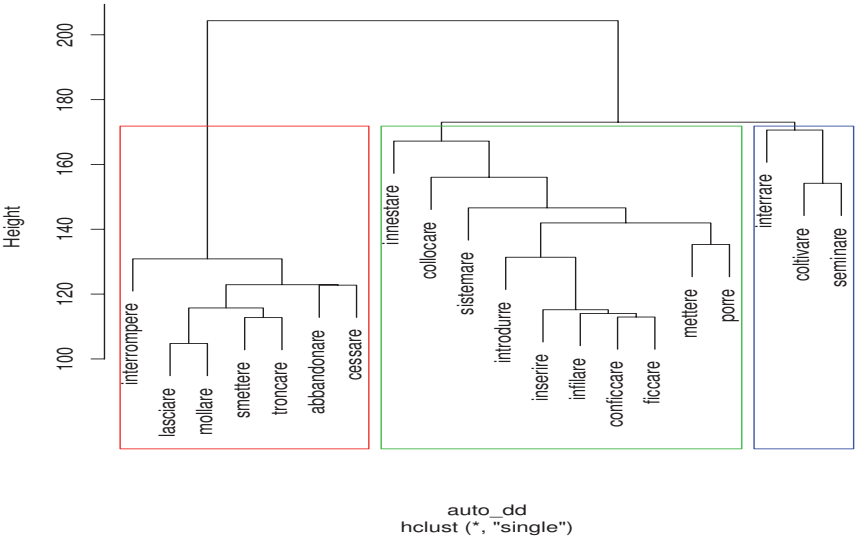

Figure 10. Hierarchical classification of verbs by single-linkage.

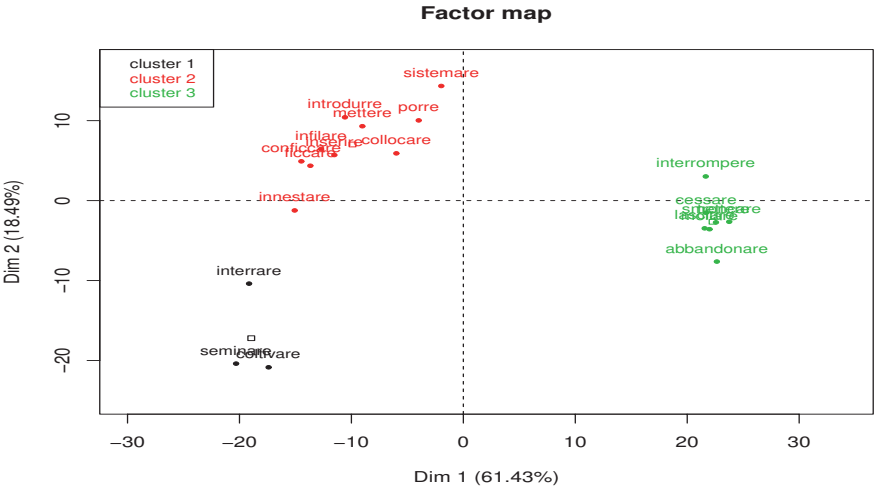

Figure 11. Three-cluster partition of verbs data by PAM and k-means.

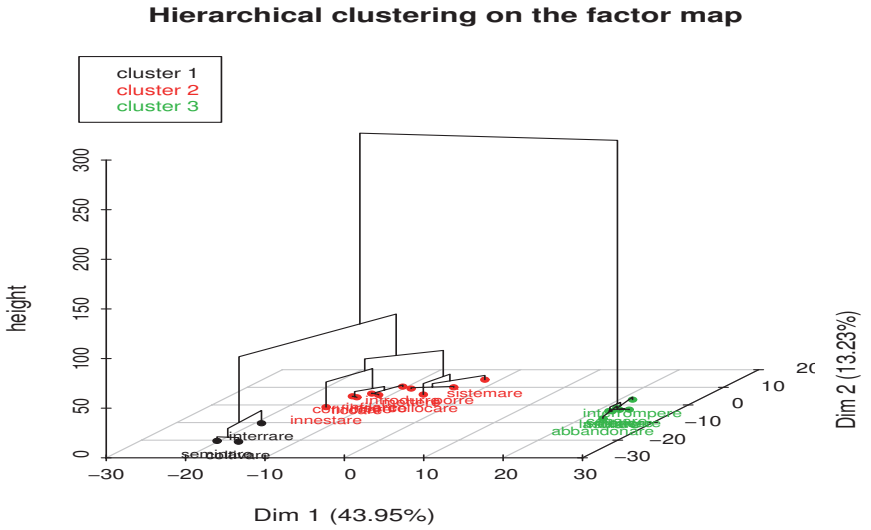

Figure 12. Representation of hierarchical clustering on the factor map.

4. Parametric approach

In this section, a statistical model to segment the items according to the ranking of perceived synonymy will be presented. The assignment of a rank to a given item entails an elicitation strategy, either referring to a perceived order or being induced from an evaluation measurement.

Landauer and Dumais (1997), in an approach they called latent semantic analysis (LSA), showed that the acquisition and comprehension of word meaning depend upon the processing and extraction of a previously unexamined kind of information (hidden in word context and past word usage), that is higher-order (or indirect) associations. It arises from the past associations that every word has with the others. But a diverse and heterogeneous collection of meanings are also context dependent (and so based upon higher-order associations). For example, some researchers emphasize the social meaning of human interactions and intentionality. Thus, the process of selection is related to subjects' experience but it is also affected by uncertainty, as it happens for any individual behaviour. This is especially true in the linguistic context, since we deal -by discrete tools (the ranks)- with feeling (here is the perceived level of synonymy) which is intrinsically continuous, but does not admit a direct measurement. This aspect is strongly related to the fuzziness of ranking procedures.

Generally, models for ranking analysis, as described by Critchlow *et al.* (1991),

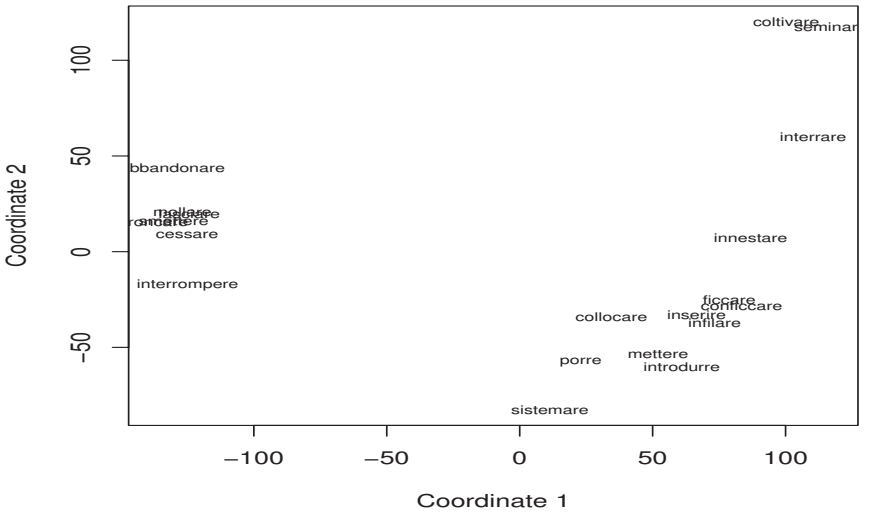

Figure 13. Non-metric Multidimensional Scaling.

Fligner and Verducci (1999) and Marden (1995), are based on order statistics, paired comparisons, distances between permutations, and stagewise decompositions of the ranking process. For the analysis of this class of models, the focus is on the simultaneous comparison of all the items with a multistage criterion, as in Kendall (1950).

Instead, in the following we adhere to a different modelling strategies by using a class of discrete mixture distribution characterized by the main component that generate an ordered choice among a list of structured alternatives. More specifically, such models allows for considering both the closeness towards the target word expressed by the respondent and also the indecision which accompanies such a choice.

4.1. Specification of CUB *models*

In order to make explicit the weight of the *uncertainty component* in a discrete model and, then, to allow for its estimation, we introduce a statistical framework denoted as CUB models (Piccolo, 2003; D'Elia and Piccolo, 2005; Iannario and Piccolo, 2012): the acronym stems from the circumstance that these models are defined as a convex ***C***ombination of ***U***niform and shifted ***B***inomial random variables. More specifically, to underline subjects' motivations which support the ordered perception of synonymy (combined with *a priori* knowledge and personal background), a mixture distribution

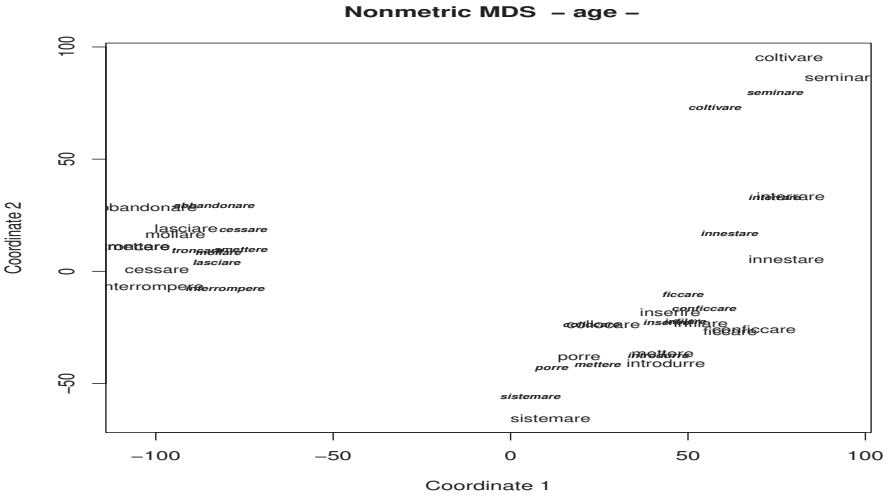

Figure 14. Non-metric Multidimensional scaling with covariate age (red: Young, black: Elderly).

of two components (*feeling* and *uncertainty*) has been introduced. In general, *feeling* is mainly related to the subjects' motivations whereas *uncertainty* mostly depends on the circumstances that surround the elicitation process.

We limit the analysis to the univariate (marginal) distributions of single items and specify and sequentially fit CUB models to univariate distributions of ranks. Of course, marginal distributions are not independent since ranks are related each others given that their sum is a constant, for a given number of items (m). However, in our experiment, $m = 20$ and so we may expect a substantial quasi-independence among the ordinal responses given by a subject. Instead, the independence of the sampled values is always preserved.

We consider the marginal distribution of the ranks expressed towards a given word as the realizations of the random variable R which assumes values over the support $\{1, 2, \ldots, m\}$ on the basis of the distribution that sampled respondents assign to the given word. Explicitly, we are considering that a marginal ranking analysis produces an indirect evaluation since people are not immediately expressing a score for the item; however, the expressed rank is an ordered evaluation as it conveys the *closeness* of the word to the given target as perceived by the subject on a graduated scale.

This random variable accounts for a personal feeling towards the item (here, the perception of synonymy considered as the perceived closeness to the given word) and an inherent uncertainty surrounding the selection of a rank. Indeed, we are interpreting each respondent choice as a weighted mixing between a meditated option and a fuzzy

decision; this weight discriminates between a full conscious respondent and a totally uncertain one (further discussion of this interpretation may be found in Iannario and Piccolo, 2010; 2012).

As a latent variable, *feeling* is the result of a judgement process which depends on several causes and can be assumed to follow a Gaussian distribution; its discrete formulation on the support $r = 1, 2, \ldots, m$ may be well characterized by the shifted Binomial random variable. Instead, uncertainty is a less defined component which turns out as a combination of partial knowledge of item, personal interest or engagement in the subject, time spent for elaborating the decision, laziness or apathy towards the topic, and so on. The worst instance of making decisions is a complete indifference (=*equipreference attitude*) and this situation is expressed as a discrete Uniform random variable U which is the result of a complete randomized mechanism where each category has a constant probability $1/m$ for any $r = 1, 2, \ldots, m$.

Notice that we are not assuming that a portion of respondents acts in a purely random manner but we consider that each person conveys a proportion of this extreme behaviour; thus, the discrete Uniform distribution is just a building block for modelling *uncertainty* in ordinal choices according to CUB models.

Formally, for a known integer $m > 3$, we define R a CUB random variable with parameters π and ξ if is characterized by the following probability distribution:

$$P_r\left(R = r\right) = \pi \left[\binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} \right] + (1-\pi) \left[\frac{1}{m} \right], \quad r = 1, 2, \ldots, m\,.$$

Uncertainty is inversely related to π whereas ξ is a direct measure of *closeness*, that is the perceived strength of synonymy of the selected word with respect to the targeted verb. The parametric space of R is the left-open unit square since $\pi \in (0, 1]$ and $\xi \in [0, 1]$ and such models has been proved identifiable for $m > 3$ (Iannario, 2010). Statistical discussion and parametric inference are fully discussed by Iannario and Piccolo (2012) whereas an effective implementation of the software in the R environment is freely available (Iannario and Piccolo, 2009).

An important feature of CUB models is the fact that -by changing parameters- the resulting shape is adequate for fitting several different empirical distributions (skewed, flat, symmetric, etc.). Moreover, if we get information on the raters' features, we can develop models linking the expressed ranks to individual covariates. In such a way, we can relate the ranks elicitation process and its components (*closeness* and *uncertainty*) to a set of individual features of the raters (age, gender, educational level, job, income, etc.).

Specifically, we first examine the CUB models on the global ranking distribution and then, we analyse the statistical interpretation of ranking expressed by different groups. Consider a dichotomous situation (Iannario, 2008) where the sample is characterized by groups G_0 and G_1, respectively. Here, we will denote by D_i a variable assuming values 0 and 1 when the i-th subject S_i, $i = 1, 2, \ldots, n$, belongs to one of the groups G_0 and G_1, respectively.

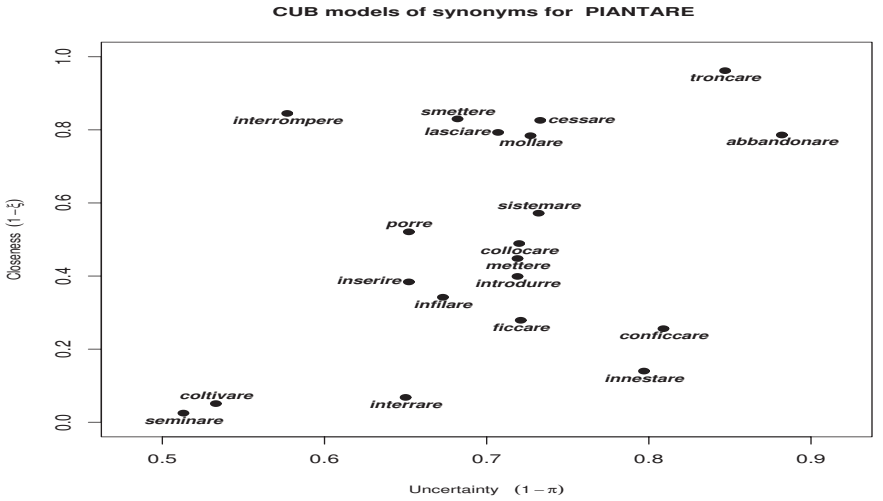

Figure 15. Visualization of CUB *models for synonymy data.*

Formally, if we suppose that this membership is relevant for explaining a different effect of the *uncertainty* and/or the *closeness* components, we specify a CUB model where the corresponding parameters are function of the dummy covariates, that is:

$$(\pi_i \mid D_i) = \frac{1}{1+\exp(-\beta_0 - \varphi D_i)}\,; \qquad (\xi_i \mid D_i) = \frac{1}{1+\exp(-\gamma_0 - \psi D_i)}\,;$$

for $i = 1, 2, \ldots, n$, where β_0 and γ_0 are level parameters, whereas φ and ψ are the parameters related to dummy effects. If $\varphi > 0$, *Uncertainty* $(G_0) >$ *Uncertainty*(G_1), and thus the corresponding π-parameters will be: $\pi_1 > \pi_0$. Instead, if $\psi > 0$, *Closeness*$(G_0) <$ *Closeness*(G_1), and the corresponding ξ-parameters will be: $\xi_1 > \xi_0$.

Interesting interpretations are also related to the expectation if we consider that mean values are related to latent components (intrinsically continuous). Specifically, we get:

$$E(R) = \frac{m+1}{2} + \pi\,(m-1)\,\left(\frac{1}{2} - \xi\right)\,;$$

thus, both parameters are involved in the specification of the expected value. As a consequence, infinite values of (π, ξ) generates the same mean value which cannot be considered as a convenient synthesis of the expressed ranks. Then, it seems more sensible to link directly the parameters to subjects' covariates as proposed in the CUB modeling paradigm.

4.2. Some empirical evidence

Analyses of CUB models for the whole sample are reported in Figure 15. If we consider the *closeness*, it shows the presence of three clusters; each set of near synonyms are under a common, coarse-grained meaning and provides a mechanism for representing a specific aspect of attitude and style. We could consider in the top of Figure the words which evoke the most scored synonymy (*agricultural* context). In the middle, we observe words related to *instrumental* factors whereas in the bottom it is possible to observe synonyms which concern *sentimental* attitude. The *uncertainty* expressed for this ranking is not so moderate and for verbs as "troncare", "abbandonare" and "conficcare", innestare" is very high (Figure 15).

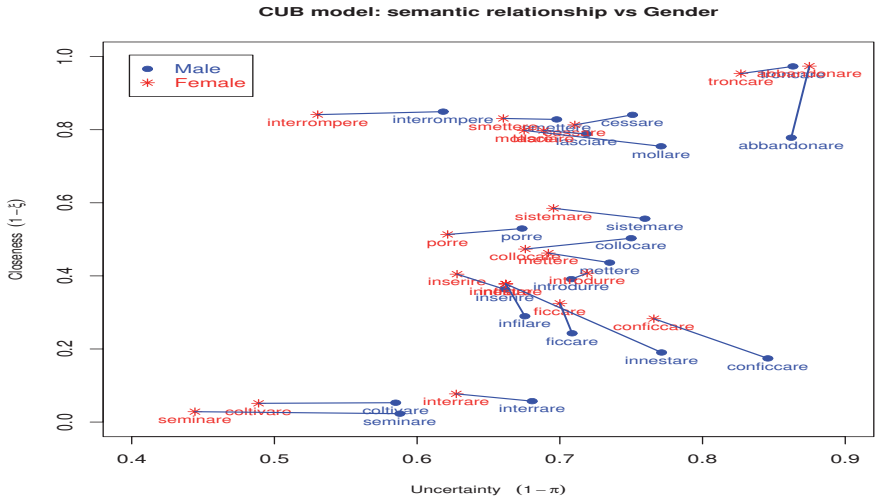

Figure 16. Visualization of CUB *models for synonymy data by Gender.*

Results related to the implementation of the CUB models with a covariate as Gender (Dummy=1 for Women) confirm the determination of the three clusters (agricultural, instrumental, sentimental) for this selected dummy (Figure 16).

Thus, it turns out that men express a greater uncertainty in the answers. Generally, in the first cluster it is not possible to emphasize particular differences in the selected arrangement; in the second, instead, some different value is present (it is preserved, nevertheless, the arrangement). These features are emphasized if we estimate CUB models for each gender as shown in Figure 17. Only for the verbs "abbandonare", "conficcare" and "innestare" we get a substantial difference in the perception of closeness with respect to Gender.

Finally, in Figure 18 we consider the *closeness* and *uncertainty* with respect to Age

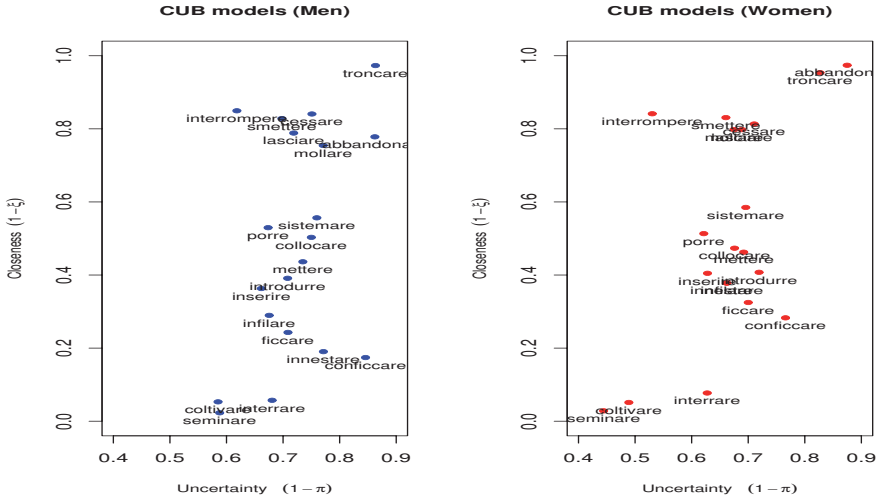

Figure 17. Visualization of CUB *models for synonymy data (left: Men, right: Women).*

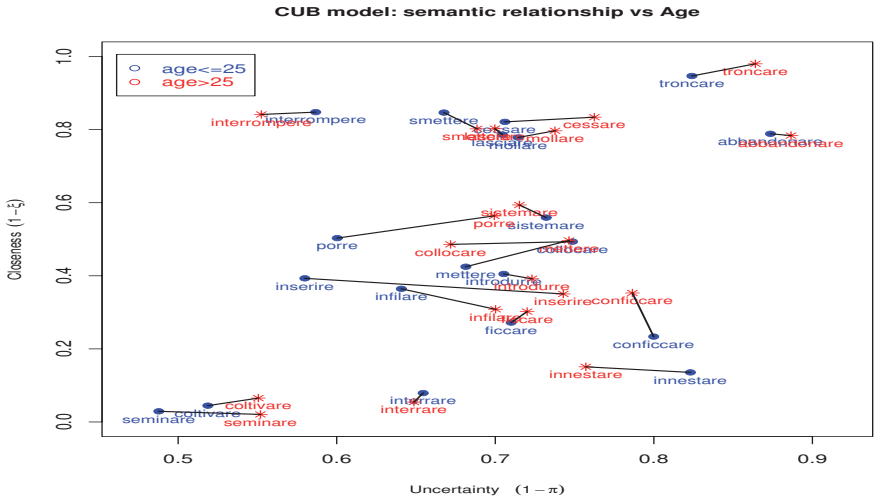

Figure 18. Visualization of CUB *models for synonymy data (by Age).*

by distinguishing people under and over 25 years. Responses are similar but in uncertainty and we observe how such difference is high for the verbs “inserire” and “porre”

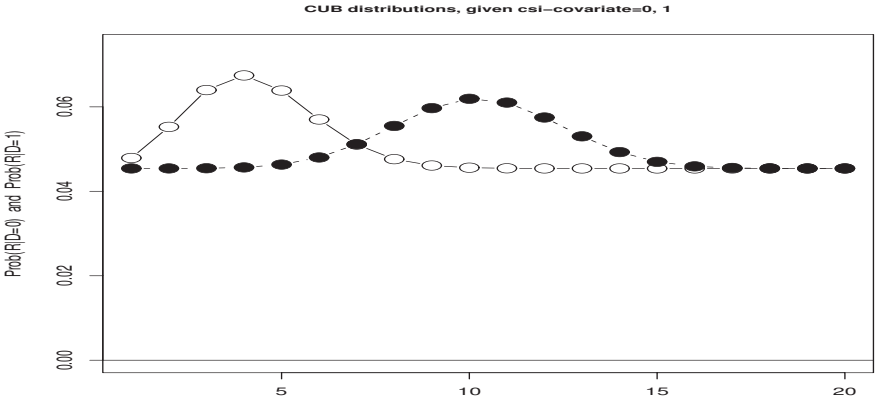

Figure 19. Estimated probability distributions of the verb "innestare" by Gender (circles are for men).

whose meaning causes greater difficulty in young respondents.

The proposed models allow for an estimation of the profiles of respondents given the selected covariates (when significant). As an instance, we choose to compare the profiles of the probability distribution of the rank of the verbs "innestare" for men and women. More precisely, we estimate a CUB model by inserting a dummy covariate for Gender as specified in the previous subsection. It turns out that Gender is not significant for explaining uncertainty but it is an important covariate for discriminating the perception of closeness among the verbs "piantare" and "innestare". The corresponding profiles are depicted in Figure 18 and confirm the separateness of the two distributions in this case.

5. Concluding remarks

The empirical analysis developed in this paper with a large data set consisting of the the ranking of synonyms of the Italian verb "piantare" has been conducted with the aim to compare nonparametric and parametric approaches for ordinal data in order to see if and when they are complimentary or opposite.

The results support common evidences and also the ability of non parametric analysis to discover a latent structure in the data allowing for a clear evidence of a semantic space of the meaning of such alternative verbs. On the other point of view, we discussed an alternative techniques for ordinal data where CUB models proved to be effective for a visualization of the verbs in the parametric space which exactly confirms the same representation obtained with the nonparametric methodologies.

Specifically, all the methods lead to three sufficiently distinct groups among syn-

onyms. The first group, with basically joins the *agricultural* meaning of "coltivare" and "seminare". The third group, which is still very compact, identifies the aspects defined as *sentimental*. Finally, the second group defined as *instrumental*, more linked to the first one and sufficiently articulated around the meaning of "posizionare" (locate), displays the synonyms as distributed in a less compact manner than the two previous ones.

As a conclusion, we strongly support the idea that both approaches are necessary and useful for the analysis of empirical data, and this general consideration is even more stringent when we are faced with ordinal data which manifest some peculiarities with respect to continuous variables.

Acknowledgements: This work has been realized with the partial support of the PRIN2008 project: "Modelli per variabili latenti basati su dati ordinali" (CUP n. E61J10000020001) at University of Naples Federico II and within FARO 2011 project. The first Author benefits of a Fulbright scholarship grant by visiting Department of Statistics and Actuarial Sciences, University of Iowa, USA.

References

Agirre E., Edmonds P. (2006), *Word Sense Disambiguation: Algorithms and Applications*, Springer, New York, Introduction, 1–28.

Anderson T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, (2^{nd} edition, 1984), J. Wiley & Sons, New York.

Bartholomew D. J. (1980), Factor Analysis for Categorical Data, *Journal of the Royal Statistical Society, Series B*, 42, 293–321.

Beh, E. J. (2008), Simple Correspondence Analysis of Nominal-Ordinal Contingency Tables, *Journal of Applied Mathematics and Decision Sciences*, 12, 1–17.

Benzécri J. P. (1973), *L'analyse des Données*, 2 Tomes, Dunod, Paris.

Camminatiello I., D'Ambra L. (2010), Visualization of the significative explicative categories using CATANOVA method and non-symmetrical correspondence analysis for evaluation of passenger satisfaction, *Journal of Applied Quantitative Methods*, 5, 64–72.

Cappelli C. (2003), Identifying word senses from synonyms: a cluster analysis approach, *Quaderni di Statistica*, 5, 105–117.

Cappelli C., D'Elia A. (2004), La percezione della sinonimia: un'analisi statistica mediante modelli per ranghi, in: Prunelle G. *et al.* (eds.): *Le poids des mots. Actes de JADT04*, Presses Universitaires de Louvain, Belgium.

Cappelli C., D'Elia A. (2006), A tree-based method for variable selection in models for ordinal data, *Quaderni di Statistica*, 8, 125–135.

Cappelli C., D'Elia A. (2011), Grouping near-synonyms of a dictionary entry: thesauri and perceptions, *Quaderni di Statistica*, 13, this issue.

Church K., Ward W., Gale P., Hanks D., Hindle, Moon, R. (1994), Lexical substitutability, in: Atkins, B.T.S. and Zampolli A. (eds.): *Computational Approaches to the Lexicon*, Oxford University Press, Oxford, 153–177.

Cox T.F. & Cox, M.A.A. (2001), *Multidimensional Scaling* , Chapman and Hall.

Critchlow D. E., Fligner M. A., Verducci J. S. (1991), Probability models on rankings, *Journal of Mathematical Psychology*, 35, 294–318.

Cruse D. A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge, UK.

D'Elia A., Piccolo D. (2005), A mixture model for preference data analysis, *Computational Statistics & Data Analysis*, 49, 917–934.

Di Marco C., Hirst G., Stede M. (1993), The semantic and stylistic differentiation of synonyms and near-synonyms, in: *AAAI Spring Symposium on Building Lexicons for Machine Translation*, Stanford, CA, 114–121.

Edmonds P., Hirst G. (2002), Near-synonymy and lexical choice, *Computational Linguistics*, 28, 105–144.

Fligner M. A., Verducci J. S., eds. (1999), *Probability Models and Statistical Analyses of Ranking Data*, Springer-Verlag, New York.

Giordan M., Diana G. (2011), A clustering Method for Categorical Ordinal data, *Communications in Statistics - Theory and Methods*, 40, 1315–1334.

Gower J. C. (1966), Some distance properties of latent roots and vector methods in multivariate analysis, *Biometrika*, 55, 315–328.

Harman H.H.(1967), *Modern Factor Analysis* (2nd ed.), Chicago University Press, Chicago.

Horst P. (1965), *Factor Analysis of Data Matrices*, Holt, Rinehart& Winston, New York.

Hotelling H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417–441; 498–520.

Iannario M. (2008), Dummy covariates in CUB models, *STATISTICA*, LXVIII, 179–200.

Iannario M. (2010), On the identifiability of a mixture model for ordinal data, *METRON*, LXVIII, 87–94.

Iannario M. and Piccolo D. (2009), A program in R for CUB models inference, Version 2.0, http://www.dipstat.unina.it/CUBmodels1/

Iannario M., Piccolo D. (2010), A New Statistical Model for the Analysis of Customer Satisfaction, *Quality Technology & Quantitative Management*, 7, 149–168.

Iannario M., Piccolo D. (2012), CUB models: Statistical methods and empirical evidence, in: Kenett R. S. and Salini S. (eds.): *Modern Analysis of Customer Surveys: with applications using R*, J. Wiley & Sons, Chichester, 231–258.

Inkpen D. (2007), A statistical model for near-synonym choice, *ACM Transactions on Speech and Language Processing*, 4, 1–17.

Inkpen D., Hirst G. (2006), Building and using a Lexical Knowledge-base of near-synonym differences, *Computational Linguistics*, 32, 223–262.

Jardine N., Sibson R. (1971), *Mathematical Taxonomy*, J. Wiley & Sons, London.

Jokinen J., McDonald J. W., Smith P. W. F. (2008), *Meaningful regression and association models for clustered ordinal data*, Methodology Working Paper M05/08, Southampton Statistical Sciences Research Institute, University of Southampton.

Jöreskog K. G., Moustaki I.(2001), Factor Analysis of Ordinal Variables. A comparison of three approaches, *Multivariate Behavioral Research*, 36, 347–387.

Kaufman L., Rousseeuw P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, J. Wiley & Sons, New York.

Kendall, M. G. (1950), Discussion on symposium on ranking methods, *Journal of the Royal Statistical Society, Series B*, 12, 189.

Kruskal J. B. (1964), Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 29, 1–27.

Landauer T. K., Dumais, S. (1997), A Solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review*, 104, 211–240.

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

Lebart L., Piron M., Morineau A. (2006), *Statistique Exploratoire Multidimensionnelle*, 4^{eme} edition, Dunod, Paris.

Manisera M., van der Kooij A. J., Dusserldorp E. (2010), identifying the component structure of satisfaction scales by nonlinear principal component analysis, *Quality Technology and Quantitative Management*, 7, 97–115.

Maravalle M.(2007), *Introduzione all'analisi dei dati con R*, Edizioni Universitarie Benedetti, L'Aquila.

Marden J. I. (1995), *Analyzing and Modelling Rank Data*, Chapman & Hall, London.

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

Pearson K. (1901), On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 1236–1238.

Piccolo D. (2003), On the moments of a mixture of Uniform and shifted Binomial random variables, *Quaderni di Statistica*, 5, 85–104.

R Development Core Team (2011), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/

Reiter E., Sripada S. (2002), Human Variation and Lexical Choice, *Computational Linguistics*, 28, 545–553.

Schoenberg I. J. (1935), Remarks to Maurice Fréchet article: "Sur la definition axionmatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert", *Annales de Mathematiques*, 36, 724–732.

Shepard R. N. (1962a), The analysis of proximities: multidimensional scaling with unknown distance function, I, *Psychometrika*, 27, 125–139.