Quaderni di Statistica
Vol. 13, 2011

Grouping near-synonyms of a dictionary entry: thesauri and perceptions

Carmela Cappelli
Department TEOMESUS, University of Naples Federico II
E-mail: carmela.cappelli@unina.it

Summary: We propose a cluster analysis approach to quantify near-synonymy relations and compare non-parametric and parametric methods. The first approach is model free since it does not assume an underlying model of lexical knowledge but it uncovers the group structure in the set of near-synonyms of a target word by comparing the list of synonyms of the given entry with those of its near-synonyms as contained into available thesauri. Then, in order to validate the results provided by the cluster analysis, a statistical model is introduced for analyzing human judgments of perceived degree of synonymy, also by a relationship with subjects' characteristics. Empirical evidence for a selected word of Italian is presented and discussed.

Keywords: Synonymy, Ordinal data, CUB models.

1. Introduction

In the last decades the application of statistical methods to various problems concerned with the analysis of textual data has attracted lot of attention: for a review, see Woods *et al.* (1986) and Lebart *et al.* (1998).

In this article, we focus on the problem of representing synonymy relations. In general, synonymy denotes equivalence of meaning but in practice it is rare to encounter absolute synonyms i.e. words with identical meaning: therefore, synonymy is rather a question of degree. Thus, thesauri actually contain *near-synonyms* i.e. words that have the same meaning but differ in lexical nuances (Inkpen and Hirst, 2006): for a given entry, a standard thesaurus provides a list of near-synonyms and it rarely offers any explanations about the shared meaning as well as about the differences in nuances between the possible choices. Indeed, discrimination between near-synonyms is a crucial aspect of mastering a language and it has relevant implications in several applications

and domains such as: Natural Language Generation (Inkpen and Hirst, 2006), translation process and writers' assistant systems (Inkpen, 2007), second language teaching and learning (Ouyang *et al.*, 2009), information retrieval, semantic tagging.

It is also to be considered that words may be ambiguous due the fact that they may have more than one meaning (polysemy). As a consequence, any study of near-synonymy relations should account for the degree of synonymy and for the presence of different meanings.

Recently, several papers have addressed the problem of representing near-synonyms: Edmond and Hirst (2002) proposed a clustered model, in which each cluster has a core denotation representing the essential meaning of the near-synonyms in the cluster and a complex internal structure accounting for semantic, stylistic and expressive differences. Inkpen and Hirst (2006), by using a simplified version of Edmonds's representation of clusters, have proposed a procedure based on a decision-list algorithm to extract the content of all entries in a dictionary of near-synonyms aiming at automatically create near-synonym representations. In a further paper, Inkpen (2007) presented a statistical method for automatic selection of the best near-synonym to be used in a given context whereas in Aminul and Inkpen (2010) the same task is addressed by means of a 5-gram model built from the Google Web 1T data set.

The work we present in this paper it is similar in spirit to the above cited papers. In particular, our goals are:

- to quantify near-synonymy by exploiting the list of near-synonyms of a target word provided by thesauri in such a way to group them by different lexical nuances;

- to evaluate our theoretical results by means of human data obtained by asking about the degree of perceived synonymy.

This goal is achieved by means of two different statistical methods. We first employ an approach which is model free and data driven in the sense that it does not assume any underlying lexical model but it creates clusters of near-synonyms of a target word by matching the brute list of its near-synonyms with those of its near-synonyms as contained into available thesauri. Indeed, the matching resorts to a binary data matrix that allows the use of cluster analysis: thus, by employing a measure of similarity among the near-synonyms that expresses analogous near-synonymy behavior with respect to the target word, the set of near-synonyms of the latter can be partitioned into groups, each group being characterized by a high degree of internal interchangeability. The groups therefore would represent different "senses" in which the target word can be meant, translated, used, etc. This approach may be defined as an objective method.

In order to validate the theoretical proposal we use human data and ask the language users/speakers to rate the near-synonyms of the target word on the basis of their perceived degree of synonymy. Subsequently, we employ a class of models (denoted as CUB models) to study the perception of synonymy, also relating it to the individual fea-

tures (e.g. age, gender, education, etc.) of the users/speakers themselves. This approach may be defined as a subjective method.

The remaining of the paper is organized as follows: in section 2 the use of the lists of near-synonyms provided by thesauri that leads to the clustering method is explained and in section 3 the results of an application to a selected word of Italian is discussed in details. Then, in section 4 the class of CUB models is introduced and applied to the synonymy of a single Italian word whereas in section 5 some empirical evidence for the selected word has been obtained to show the usefulness to make reference to subjects' characteristics. Some concluding remarks follow in section 7.

2. Cluster analysis for grouping near-synonyms

Since our goal is to uncover differences among the near-synonyms of a target word, cluster analysis appears a natural candidate approach because it reveals a group structure in the data, where the groups contrast with each other and are internally homogeneous.

Any statistical textual analysis requires qualitative information to be turned into quantitative data: thus, we need a *working hypothesis*. Specifically, we assume that any word has a *semantic space* spanned by its near-synonyms[1] so that the problem reduces to describe and partition this space. Note that the same hypothesis has been formulated by Ploux (2003) but, in that case, near-synonymy was studied in the context of translation, that is aiming at matching French-English dictionaries, by means of correspondence analysis applied to cliques of near-synonyms.

Based on the above mentioned hypothesis, let w be a target word and define the extended set $S_w = [s_1, ..., s_j, ..., s_m]$ of its near-synonyms as the set of all possible near-synonyms of w collected by the available sources (manually compiled and electronic thesauri). For any s_j, let us consider the corresponding expanded set of near-synonyms S_{s_j}. Comparing the set S_w with each set S_{s_j}, $j = 1, \ldots, m$, yields to building a $m \times (m+1)$ data matrix whose rows are the elements in S_w and the columns are the same items in S_w plus w itself. In particular, in the matrix, by row, each near-synonym s_j of w is represented by a sequence of 1s and 0s according to whether each term in S_w and the same word w is present or absent in S_{s_j}. In other words the synonymy behavior is studied conditioned to the word of interest, that is, to the reference set S_w. The reason why the entry term w is also considered among the variables is that, at least in the Italian dictionaries of synonyms, we have found that synonyms are not always reciprocal. The resulting data matrix has the form reported in Table 1 where $\delta_{jj'}$ equals 1 if the j'-th near-synonym of the entry term is also a near-synonym of s_j and equals 0 otherwise.

Based upon this binary data matrix, a similarity measure can be specified. In general, measures proposed in the literature for binary data differs for the inclusions of conjoint absence of attribute. It is our belief that in this context the conjoint absence (hence-

[1] Of course, this hypothesis should not to be considered exhaustive. Alternative hypotheses could be formulated, for example by involving also the antonyms.

Table 1. Structure of the binary data matrix

	w	s_1	...	...	$s_{j'}$	...	s_m
s_1					$\vdots$		
$\vdots$					$\vdots$		
s_j	...	...	...	...	$\delta_{jj'}$	...	...
$\vdots$					$\vdots$		
s_m					$\vdots$		

forth denoted as CA) of an attribute is indicative of similar synonymy behavior because similarity is evaluated conditioned to the target word and in this respect it is qualified by what a term might mean (presence) as well as by what it might 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 CP plus conjoint absence CA) to the total:

$$SM = \frac{CP + CA}{m + 1}.$$

This measure ranges from 0 to 1, being 0 if there are no matches and 1 if there are no mismatches.

This measure is computed for any pairs of items and yields to a similarity matrix on which 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. Then, the inspection of the hierarchy allows for the identification of:

- *broad senses*, in which the semantic space of the given term can be partitioned, represented by nodes placed on top of the dendrogram;
- *patterns*, that is finer and finer senses corresponding to nested clusters generated at lower levels in the hierarchy;
- *outliers*, that is senses far from the others or extremely generic which are connected on top of the hierarchy.

This proposal applied to all entries in the dictionaries of near-synonyms would resort to an electronic dictionary in which, for any entry, the list of near-synonyms is not fixed but it can vary by imposing different thresholds on the index of the hierarchy or choosing alternative similarity measures or aggregation criteria. In other words, it would represent an additional resource for word sense disambiguation.

3. The case of the noun Scolaro

In this section we present and discuss the results of the application of the proposed clustering approach to the case of the Italian word *Scolaro*, whose extended set of near-synonyms is: {*Allievo, Alunno, Discente, Discepolo, Educando, Seguace, Studente*}; for the complete list of thesauri considered for this research, see Cappelli (2003). Table 2 reports the binary data matrix built according to the matching procedure described in Section 2.

Table 2. Binary data matrix for the noun "Scolaro"

	Scolaro	*Allievo*	*Alunno*	*Discente*	*Discepolo*	*Educando*	*Seguace*	*Studente*
Allievo	1	1	1	1	1	0	1	1
Alunno	1	1	1	1	1	0	1	1
Discente	1	1	1	1	1	0	0	1
Discepolo	1	1	1	0	1	0	1	0
Educando	1	0	0	0	0	1	0	1
Seguace	1	1	0	0	1	0	1	0
Studente	1	1	1	1	1	1	0	1

A main issue in any hierarchical clustering method is the choice of the aggregation criterion. Since no one criterion is best or can be unconditionally recommended, as a default procedure we have considered several ones, namely: the *single linkage*, the *complete linkage* and the *average linkage*. Figure 1 depicts the dendrogram for the near-synonyms of *Scolaro* obtained using the average linkage. It is worth noticing that all criteria gave identical results on these data and reveal the same group structure.

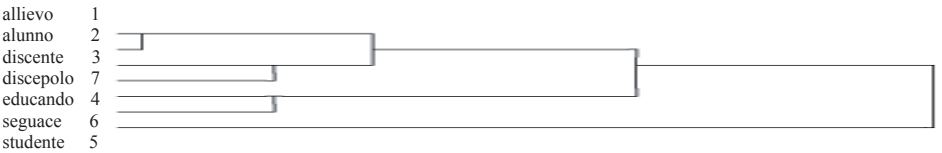

Figure 1. Dendrogram for the near-synonyms of noun Scolaro

The dendrogram is to be interpreted as follows: the high levels in the hierarchy represent large clusters that identify main (broader) senses whereas aggregations at lower

levels represent specific senses. We see that three groups stand out, corresponding to three main senses of noun *Scolaro*:

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 living in a college.

Notice that sense 1, that can be generically meant as "learner", is split to a lower level of the dendrogram into two more specific senses: "pupil" {*Allievo, Alunno*} and "student" {*Discente, Studente*}. In particular, {*Allievo, Alunno*} are the first synonyms of *Scolaro* to be merged, that is, they represent the most interchangeable words between them and with respect to *Scolaro*.

We also observe that the synonym *Educando* appears far away from all the other, identifying always a separate sense.

In order to evaluate the global fit of the hierarchy to the data, we have considered the *cophenic correlation coefficient* that it is computed between the $m(m-1)/2$ values in the lower half of the similarity matrix and the corresponding values in the so called cophenic matrix. This matrix has been built by considering, for any pair of objects, the first level at which the two objects are grouped in the hierarchy. The value of this measure is 0.879 and confirms that the data have a strong hierarchical structure.

4. A statistical model for the perception of synonymy

In this section our attention switches over the human judgement of near-synonymy. At this aim we present a statistical model for the analysis of perceived synonymy; then, in the following section we verify whether the information content provided by thesauri and coded by the cluster analysis approach, as discussed in Sections 2 and 3, matches the perception of the language users/speakers.

Let us assume that n raters are asked to elaborate a ranking of given m elements in the extended set of near-synonyms S_w, by assigning rank $R = 1$ to the the synonym that is felt as the most similar with respect to w, rank $R = 2$ to the "second best" synonym, and so on. Thus, for each element s_j $(j = 1, 2, ..., m)$ in S_w we get a vector of observed ranks $(r_1, r_2, ..., r_n)$, that represent the degree of perceived synonymy between s_j and w, according to the opinion of the n raters. From a statistical point of view, for a given m, we observe that the information contained in the individual ranks $(r_1, r_2, ..., r_n)$ is strictly equivalent to the distribution of frequencies $(n_1, n_2, \ldots, n_m)$, where n_j is the number of respondents who locates the selected near-synonymy at the j-th position, for $j = 1, 2, \ldots, m$.

A statistical model aiming at adequately describing the process underlying the elicitation of a perceived-synonymy ranking involves (Cappelli and D'Elia, 2004):

- an *evaluation component*, by means of which each rater assesses the strength of perceived similarity between s_j and w;

- an *uncertainty component*, due to both the semantic fuzzyness of w and subjectiveness (personal impressions, temporary emotions, local conditions surrounding the choices, boredom and lazyness, and so on).

The *evaluation component*, being related to a judgement feeling, is intrinsically continuous. Nevertheless, since we observe a discrete realization (the rank r assigned to each s_j), we propose to represent it by means of a shifted Binomial random variable, that allows for a mapping on the discrete support $(1, 2, ..., m)$ and it is, in some extent, related to the pairwise comparison criterion of choice (D'Elia, 2000), widely discussed in the psychometric literature.

The *uncertainty component* can be represented by the Uniform discrete random variable, since it is known that this distribution maximizes the entropy among all the discrete distributions with finite support. Thus, it is the less informative option since it describes a completely random choice.

Finally, both components are adequately weighted by a quantity that measures the *propensity* of the respondent to adhere to a meditated or a random choice (Iannario and Piccolo, 2010).

Then, for representing the process that leads to a (near-synonymy) ranking, we introduce a class of discrete mixture models by means of a *C*ombination of a *U*niform and a shifted *B*inomial random variable (henceforth, denoted as CUB model[2]).

Then, we say that R is generated by a CUB random variable if:

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

where $\pi \in (0, 1]; \quad \xi \in [0, 1]$.

The parameters of the CUB model are easily interpretable: the quantity $(1 - \pi)$, that is the weight of the Uniform component in the mixture, is a measure of uncertainty due to the fuzziness of a given word whereas the parameter ξ is directly related to the probability of perceiving s_j as the strongest near-synonym of w. As a consequence, the estimate of ξ results in a *measure of perceived synonymy*: the higher ξ, the stronger the synonymy between s_j and w is felt by the rater. Of course, we can consider the estimate of ξ also as an *interchangeability measure* between s_j and w, derived from the empirical evidence.

An important feature of CUB models is their easy and appealing visualization of the estimated probability distributions in the parametric space by a single point: in this way,

[2] These models have been firstly proposed by Piccolo (2003) and applied to rank data by D'Elia and Piccolo (2005). A noticeable extension with subjects' and objects covariates has been obtained by Piccolo and D'Elia (2008). Identifiability of CUB models has been proved by Iannario (2010) and more recent extensions are discussed in Iannario (2012), Iannario and Piccolo (2012).

several concepts as closeness, similarity, significance and groups of near-synonyms are greatly simplified[3].

5. Empirical evidence of the and perception of near-synonymy

In order to study the perception of the synonymy with respect to the word *Scolaro*, a large sample survey has been organized during 2003 and 654 questionnaires have been validated. Respondents were asked to rank the corresponding 7 near-synonyms with respect to the noun *Scolaro*. These were presented in alphabetical order, and the respondents were asked to assign rank from 1 to 7, on the basis of the strength of perceived synonymy; no ties were allowed. In addition, based on sociolinguistic practice, for each respondent we have considered as relevant covariates: gender, age, and some proxies of socio-cultural status such as education, frequency of reading, use of Internet, frequency of travelling, other languages spoken, and so on.

Table 3. Perceived synonymy with respect to noun Scolaro.

Synonyms	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}$	$Var(r)$
Alunno	0.924	*0.005*	0.955	*0.012*	1.561	0.802
Studente	0.817	*0.007*	0.917	*0.019*	2.249	1.349
Allievo	0.712	*0.007*	0.990	*0.011*	2.745	0.823
Educando	0.376	*0.008*	0.963	*0.022*	4.720	1.409
Discepolo	0.286	*0.008*	0.945	*0.020*	5.223	1.397
Discente	0.237	*0.009*	0.824	*0.028*	5.372	1.995
Seguace	0.125	*0.006*	0.941	*0.015*	6.130	1.202

In the first instance, for each of the near-synonyms of *Scolaro* the estimates of the parameters π and ξ of CUB model were obtained. Table 3 reports these estimates (with their standard errors) ordered by the values of $\hat{\xi}$, that is on the basis of the strength of perceived synonymy; also the sample mean $\bar{r}$ and the variance of the observed ranks are displayed.

It can be noticed that we get at least three groups of synonyms. The strong synonyms ($\hat{\xi} > 0.7$): $\{Alunno, Studente, Allievo\}$ that share the meaning of *someone who learns knowledge and/or is enrolled in educational program*; among these, *Alunno* is perceived as highly interchangeable ($\hat{\xi} = 0.924$) with the target word *Scolaro*. Then, we have the weak synonyms: {*Discepolo, Discente, Seguace*}; among these, the last

[3] A note of caution: since ranks are not independent (this translates into a sparseness effect of the corresponding points along the ξ axis), this property must be taken into account when interpreting and commenting the graphical output of CUB models. However, the relationship among the words may be usefully displayed and interpreted. For more specific modelling approaches to rank data, see Fligner and Verducci (1999) and Marden (1995).

three ($\hat{\xi} < 0.3$) share the meaning of *someone who takes up knowledge or beliefs from a "master"* and they are felt as very weak synonyms. Finally, the synonym *Educando* ($\hat{\xi} = 0.376$) makes a group by his own, with the meaning of *young person living in a college*. As to the uncertainty component, measured by $(1 - \hat{\pi})$, its value is homogeneously very low for all the near-synonyms except for the word *Discente* that manifests a level of indecision almost twice as much as the others.

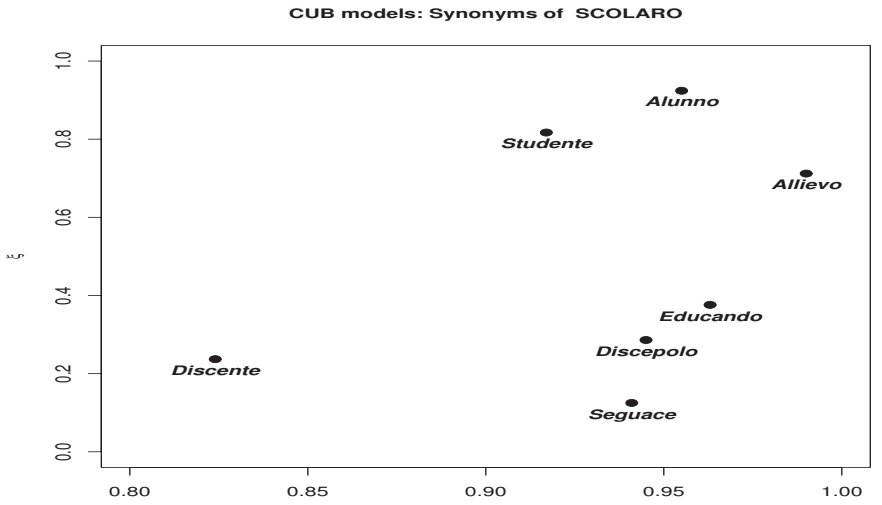

Figure 2. Visualization of CUB *models for the near-synonyms of noun Scolaro*

If we represent the estimated CUB models in the parametric space such a pattern is even clearer as shown in Figure 2: notice how the groups of near-synonyms are well identified and also the effect of working with rank data that induces a scattering of points along the ξ axis. Finally, it is evident the anomalous position of the word *Discente* that is accompanied by a larger uncertainty.

By comparing these results with those obtained in Section 3 (by using the cluster analysis approach on the thesauri), it emerges that the identification of three clusters of synonyms with different meanings is consistent with the perception of the users.

The only exception regards the synonym *Discente*: in fact, using the theasuri-based approach *Discente* is clustered together with *Allievo, Alunno* and *Studente* whereas in the perception-based approach it appears to be felt as a weak synonym together with *Discepolo* and *Seguace*. In fact, referring to the observed distribution of ranks assigned to *Discente*, we notice that most of the sample did not recognize it as a synonym of the word *Scolaro*.

6. The role of subjects' covariates in the perception of the synonymy

The CUB model allows also for the introduction of raters' covariates, aiming at relating features of the raters both to the perception and to the uncertainty contained in their choices. If we let $\boldsymbol{T}$ be the $n\times\ (v+1)$ matrix of v covariates of the n raters, whose generic row is $\boldsymbol{t}_i = (1, t_{i1}, ..., t_{iv})^{'}$, $i = 1, 2, ..., n$, then we may link the individual rater's covariates to the parameters characterizing the CUB model, by means of the logistic function, as follows:

$$(\pi_i \mid \boldsymbol{y}_i) = \frac{1}{1+\exp(-\boldsymbol{y}_i\boldsymbol{\beta})}; \qquad (\xi_i \mid \boldsymbol{w}_i) = \frac{1}{1+\exp(-\boldsymbol{w}_i\boldsymbol{\gamma})}; \quad i = 1, 2, \ldots, n\,,$$

where $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the covariates of the i-th subject extracted by $\boldsymbol{T}$ (selected according to their significance) and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the coefficient vectors of the covariates. Notice that such covariates may be coincident, partially or totally different.

By means of the model with covariates, both the uncertainty/fuzzyness component (inversely characterized by the π parameter) and the strength of perceived synonymy (directly represented by ξ) may be related to the individual features of the raters. Thus, we can analyze the role of the raters' covariates in the perception of the synonymy, and the effect they play in the way of "feeling" a word. Finally, it is possible to study the expected perception of synonymy for different profiles of users/speakers.

To further investigate the synonymy of the word *Scolaro*, we have applied CUB models with respondents' covariates to analyse the perception of the synonymy between *Discente*, *Studente* and the target noun *Scolaro*.

With reference to the near-synonymy between *Studente* and *Scolaro*, with a backward selection criterion, we found that the following covariates are significant for explaining the perception such a synonymy in terms of ξ (strength of synonymy): to have a University Degree (UnivDeg=0, Not; UnivDeg=1, Yes) and to be a Frequent Reader (RegRead=0, Not; RegRead=1, Yes). Both of them have a significant positive effect on perceiving *Discente* as a near-synonym of *Scolaro* (Table 4, left side).

It clearly appears that there exist a different perception of the synonymy between *Discente* and *Scolaro*, depending on the education level of the respondent and on him/her being a frequent reader; both aspects refer to a cultural background of the respondent. Indeed, more educated people who are also frequent readers feel more strongly ($\hat{\xi} = 0.669$) the similarity between *Discente* and *Scolaro*, opposed to less educated and not reader ones, that perceive *Discente* as a very weak synonym ($\hat{\xi} = 0.207$) of *Scolaro*. In this case study, the expected rank $E(R)$ assigned to *Discente* by different profiles of respondents is consistent with this finding (Table 4, right side).

This evidence, mainly due to the fact that *Discente* is somewhat a sophisticated word of Italian, is confirmed also in Figure 3, that displays the estimated probability distributions for the two extreme profiles (Not-Not= $[0, 0]$ vs. Yes-Yes= $[1, 1]$) of respondents (the plots of discrete distributions are connected to emphasize their different location). We see that respondents characterized by an higher cultural profile tends to perceive

Table 4. CUB *model estimates and expected perceived synonymy (word "Discente")*

Covariate	$\hat{\gamma}$	*e.s.*
Constant	-1.338	*0.047*
UnivDeg	1.021	*0.107*
RegRead	0.261	*0.123*
$\hat{\pi}$	0.854	*0.014*

Profiles			
UnivDeg	*RegRead*	$\hat{\xi}$	E(R)
Not	Not	0.207	5.497
Not	Yes	0.254	5.260
Yes	Not	0.421	4.403
Yes	Yes	0.669	3.134

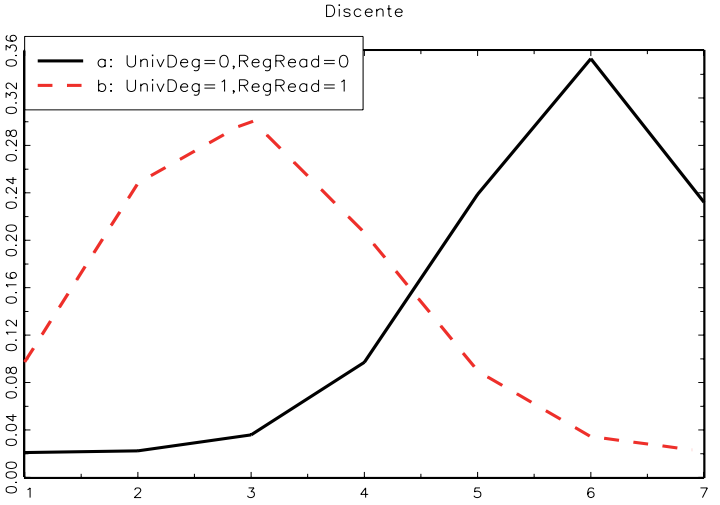

Figure 3. Probability distributions of degree of near-synonymy of the word "Discente"

more the synonymy between *Discente* and *Scolaro*, being the respective estimated distribution displaced on lower values of the ranks. In further analyses, we found that also the covariates *Age* and *Site of response* ($= 1$, University lecture hall, $= 0$, otherwise) are marginally significant covariates for explaining the perceived degree of synonymy. They confirm the importance of the cultural background on the perceived synonymy of the word "Discente".

As a second case study, we consider the perceived synonymy between *Studente* and *Scolaro*. Thus, in Table 5, we present the estimated parameters (asymptotic standard errors of parameters are in parentheses) and the corresponding log-likelihood functions of CUB models obtained by a stepwise procedure in order to achieve a better result. Notice that *Gender* is a dichotomous variable ($= 1$ for women, $= 0$ for men) whereas *Age* is a continuous covariate obtained by log-transforming the years of the respondent and then subtracting the average.

Table 5. Stepwise CUB *model estimates for the perceived synonymy of "Studente"*

Covariate	Uncertainty parameters	Synonymy parameters	log-likelihood
No Covariate	$\hat{\pi} = 0.917$ *(0.019)*	$\hat{\xi} = 0.817$ *(0.007)*	-936.89
Constant	$\hat{\beta}_0 = 2.638$ *(0.317)*	$\hat{\xi} = 0.820$ *(0.007)*	-928.86
Age	$\hat{\beta}_1 = -2.295$ *(0.599)*		
Constant	$\hat{\pi} = 0.931$ *(0.018)*	$\hat{\gamma}_0 = 1.428$ *(0.048)*	-926.10
Age		$\hat{\gamma}_1 = -0.551$ *(0.114)*	
Constant	$\hat{\pi} = 0.918$ *(0.019)*	$\hat{\gamma}_0 = 1.638$ *(0.073)*	-933.00
Gender		$\hat{\gamma}_1 = -0.263$ *(0.095)*	
Constant	$\hat{\pi} = 0.929$ *(0.018)*	$\hat{\gamma}_0 = 1.627$ *(0.073)*	-922.14
Age		$\hat{\gamma}_1 = -0.548$ *(0.114)*	
Gender		$\hat{\gamma}_2 = -0.262$ *(0.094)*	
Constant	$\hat{\beta}_0 = 2.619$ *(0.296)*	$\hat{\gamma}_0 = 1.626$ *(0.072)*	-919.94
Age	$\hat{\beta}_1 = -1.438$ *(0.657)*	$\hat{\gamma}_1 = -0.455$ *(0.128)*	
Gender		$\hat{\gamma}_2 = -0.253$ *(0.094)*	

Given the information set available on the sample data, the last model is the best one and it represents a probability measure of the perceived synonymy of *Studente* with respect to *Scolaro* according to the following probability distribution:

$$Pr\left(R = r_i \mid Age_i, Gender_i\right) = \pi_i \binom{6}{r_i - 1} (1-\xi_i)^{r_i - 1} \xi_i^{7-r_i} + \frac{1-\pi_i}{7},$$

where, for $i = 1, 2, \ldots, n$,

$$\pi_i = \frac{1}{1 + e^{-2.619 - 1.438\, Age_i}}, \qquad \xi_i = \frac{1}{1 + e^{-1.626 + 0.455\, Age_i + 0.253\, Gender_i}}.$$

As a consequence, uncertainty is conditioned by the *Age* of the respondent (elderly are more uncertain) whereas the perceived synonymy depends on both *Age* and *Gender* (young and men have higher perception than elderly and women, respectively).

An important feature of CUB models is the possibility to make visible in a single plot several aspects obtained after the modelling step on the ordinal data. Thus, in Figure 4 (left panel) the behaviour of the parameters is represented in function of a varying *Age* and also by *Gender* (for the parameter ξ). This circumstance causes the CUB models to change with both covariates and we can dynamically visualize them in the parameter space, as shown in Figure 4 (right panel). This plot immediately confirms the previous evidence.

7. Concluding remarks

We have proposed a cluster analysis approach that exploits the information provided by thesauri in order to get a partition of the semantic space of a target word spanned by its near-synonyms. The application has shown that this approach is promising in

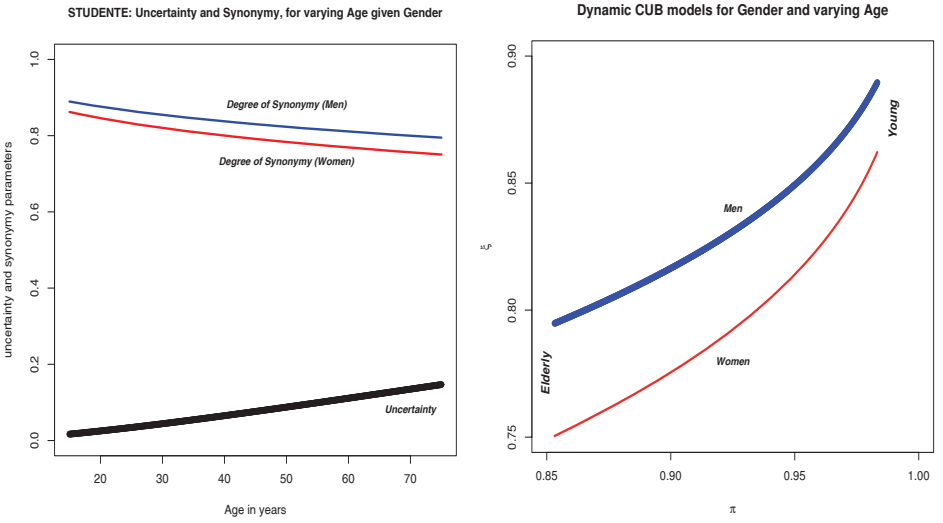

Figure 4. Dynamic representation of estimated parameter and CUB *models*

grouping near-synonyms according to the degree of synonymy with respect to the target word, also accounting for its different meaning.

In a bigger project, the proposal applied to all entries in dictionary of near-synonyms could provide an additional resource to which turn to for choosing the right word and for sense disambiguation.

Moreover, in order to get an empirical validation of the proposal, we have employed human data by introducing a statistical model for ordinal variable that allows for the analysis of the role of individual covariates (gender, education level, etc.) on the way a word is ”felt” by its users/speakers. At this regard, the results show that there is a satisfactory consistency between the findings of the thesauri-based cluster analysis approach and the human data.

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. The author would like to thank dr. Angela D’Elia for providing the initial stimulus and opportunity to develop this research.

References

Aminul, I., Inkpen, D. (2010), Near-synonym choice using a 5-gram language model, *Research in Computing Science*, 46, 41–52.

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.

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

D'Elia A. (2000), Il meccanismo dei confronti appaiati nella modellistica per graduatorie: sviluppi statistici ed aspetti critici, *Quaderni di Statistica*, 2, 173–203.

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

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

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

Iannario M. (2012), Modelling *shelter* choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, 21, 1–22.

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.

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

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

Ouyang, S., Hong Gao, H., Ngee Koh, S. (2009), Developing a computer facilitated tool for acquiring near-synonyms in Chinese and English, *Proceedings of the Eighth International Conference on Computational Semantics*, 316–319.

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

Piccolo, D., D'Elia, A. (2008), A new approach for modelling consumers' preferences, *Food Quality and Preference*, 19, 247–259.

Ploux, S. (2003), A model for matching semantic maps between languages, *Computational Linguistics*, 5, 27–48.

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

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