Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition

REMO BINDI, NICOLETTA CALZOLARI, MONICA MONACHINI, VITO PIRRELLI, and ANTONIO ZAMPOLLI
Istituto di Linguistica Computazionale del CNR Pisa, Dipartimento di Linguistica dell'Università di Pisa

Abstract

An attempt to integrate different techniques and various perspectives on lexical knowledge acquisition from text corpora is illustrated. In this program we use three distinct methodologies to handle text data, summarized as follows:

(1) Simple and traditional stochastic techniques working on pairs of words.
(2) A lexicographic approach guided by the techniques mentioned in Section 1, aiming at a formal description of sense disambiguation in terms of rules.
(3) More complex and sophisticated statistical methods working on sets of words (possibly belonging to the same semantic field), which allow us to gain a new perspective on the problem of sense disambiguation.

The three approaches are complementary to each other and can be contextually used.

The overall objective of our work is to try to integrate data and information coming from different sources, i.e. machine-readable dictionaries, text corpora, linguists' or lexicographers' knowledge, within a computational lexicon. We stress the necessity of convergence of (1) lexical and textual projects, (2) computational and traditional lexicography, and (3) statistical and rule based approaches.

1. Introduction

1.1. *The Need for Corpora in NLP*

Corpora of written and spoken language are an essential primordial resource for any NLP projects aimed at real application. It is our contention that the creation of adequate corpora is the most urgent need for the development of language engineering.

Corpora are essential sources of linguistic information. If an NLP system is to process successfully a given language for a given purpose, it must be based on the evidence of how language is really used. The analysis of corpora (i.e. representative collections of texts in machine-readable form) is the main source of obtaining this evidence. As such it is irreplaceable.

Preliminary work carried out in the context of NLP applications (in particular speech, information, retrieval, dialogue, translation) has clearly shown that it is necessary to take advantage of the pertinent characteristics and properties of the various 'sublanguages'. By sublanguages we mean different uses of the same language in different communicative contexts, with different speakers, for different communicative purposes; they present relevant differences in the range and distribution of several classes of linguistic phenomena. If the various sublanguages are objectively identified, and their pertinent features adequately described, their differences and their specific properties can be usefully exploited to reduce the actual range of computer-intractable linguistic situations, to improve the economic performance of the system, to increase the acceptability of the products, to widen up the range of effectively feasible NLP practical industrial applications.

Corpora analysis is the only tool known that is able to identify and describe sublanguages. Corpora are also the only possible source of data on the statistical properties of the various elements of languages and sublanguages: their distribution, frequency of letters, phonemes, words, categories, structures, co-occurrences, and their sequences and relations. The most successful recent NLP practical systems are heavily based on statistical evidence and methods. The integration of qualitative and quantitative methods is one of the most promising approaches to NLP, recognized both in the fields of computational linguistics and artificial intelligence. This applies for both language analysis and language generation and also in connection with the development of parallel computing and connectionistic processes.

Corpora are even more necessary for contrastive descriptions of different languages. Any multilingual application, must take into account correspondences and differences between languages, analysed on the basis of the evidence provided by multilingual corpora and by multilingual access to related monolingual corpora.

Last but not least, carefully designed representative, stratified, and classified language corpora are essential to the creation of methods and tools for the evaluation of NLP techniques, approaches, components, systems, and performances. It is a recognized fact that the nearly total lack of objective informative evaluation criteria is a major restrictive element and is one of the reasons for the hesitation of the major industries in committing themselves completely to the development of systems for which it is necessary to incorporate NLP components.

1.2. *Corpora and Computational Lexica*

In this paper, attention will be focused on a specific aspect of this comprehensive framework; namely, how corpora can be used to enhance the structure of a

Correspondence: Dr R. Bindi, Institute for Computational Linguistics, Via della Faggiola 32, 56100 Pisa, Italy.

broad-coverage computational lexicon. From this perspective, a preliminary assumption of our research work (a point which has been carefully investigated and which has gathered considerable support over the last few years) is that machine-readable dictionaries, in spite of some difficulties, have proven to be an invaluably rich source of linguistic information and a suitable basis for building up computational lexica. Nevertheless, machine-readable dictionaries do not contain all the evidence we would like to avail ourselves for eventual integration into a computational lexicon. We therefore turned to corpora as a complementary viable means of gathering lexical-driven evidence. By scouring large text corpora we aim at using texts as another source of lexical knowledge, our ultimate purpose being to integrate data thus extracted into (1) computational lexica for NLP, and (2) traditional dictionaries. It is our contention that the same information can profitably be used both for computational and traditional lexicographers. Moreover, this project fits in extremely well with our objective of devising a multifunctional computational lexicon, which can meet the needs of both 'traditional' users and NLP applications.

Our method of handling text data can be outlined as follows:

(1) Use of simple and traditional stochastic techniques which work on pairs of words.
(2) Use of a lexicographic approach guided by the techniques mentioned above (Section 1), aiming at a formal description of sense disambiguation in terms of rules.
(3) Use of more complex and sophisticated statistical methods which work on sets of words (possibly belonging to the same semantic field), allowing us to gain a new perspective on the problem of sense disambiguation.

The three stages above are complementary to each other, and can be used contextually. In the near future we intend to fully exploit this combined methodology. Existing lexical entries, whose information is elicited by merging data extracted from several machine-readable dictionaries (this phase of the work is currently being developed within the ACQUILEX ESPRIT Project, see Boguraev *et al.*, 1988) will be enriched with types of information coming from texts analysed according to the strategy outlined below. Above all, we wish to stress the necessity of convergence of (1) lexical and textual projects, (2) computational and traditional lexicography, and (3) statistical and rule-based approaches.

1.3 *Rule Writing and Use of Statistical Tools: Methodological Qualifications*

Acquisition of lexical information from large text corpora through numerical/statistical processing, is a viable tool of analysis for rule writing in at least two senses: (a) it helps in collecting the information one needs in the first stage of preliminary data-tapping, making sure one is not leaving things out (completeness); (b) it helps to check whether generalizations are being made in the right direction since sporadic phenomena cannot be treated on an equal footing with more frequent phenomena and hence more powerful generalizations must be given priority within a suitable rule-writing strategy.

The above methodology creates further questions for investigation. In the final part of our paper we will try to answer the following question: are rules the only means of 'harnessing' linguistic knowledge?

One research area in the study of lexical computation which lends itself reluctantly to any form of rule-driven approach is the identification and classification of word-senses.

The 'fuzzy edges' of word senses can hardly be accounted for by axiomatic/deductive strategies. The so-called 'check-list' theory (Fillmore, 1975), according to which word senses possess 'criterial attributes' which we can check off, one by one, to see whether a given word sense applies in a certain situation or not, may work for words like *bachelor* or *square*, but leaves us in the lurch with 'slippery' word senses like *red* or *tiger*. Most words in our day-to-day usage are unfortunately of the second, slippery type. So-called *terms*, that is, words set by firm, carefully specified boundaries, may play a prominent role in technical texts, but cannot be realistically viewed as the more common case of wording requiring the lexical competence of the native speaker.

A large variety of differing word types do not lend themselves to explanation along the lines of clear principles of individuation for word-meanings. For example, an adjective like *open*, which usually conjures up the picture of a door which lets 'air, light, things or people ... pass through' (COBUILD), takes on a different meaning when accompanied with words such as *bottle, box, letter, parcel*, etc. It means *candid* when predicated of a human being, it means *unbounded* when one is talking about sets of items, and so on and so forth. Unlike 'cup-like' objects which are identified with different names depending on the situational context they are put in (as proven in Labov's experiment), different uses of the same word appear to be stretched even beyond the limit either of an arguably shared function (i.e. all Labov's things contain something), or of a common set of shape-features (circularity, concavity, etc.). Moreover, unlike Labov's *bowls–cups–mugs–vases*, word senses do not always appear to cover a continuum of features varying smoothly and uniformly. In many cases we actually come across apparently inexplicable gaps, while in others we come across overlapping meaning areas.

In the following paper, particularly in the final part, we will illustrate various methods to characterize different 'word senses' without assuming that the word sense is 'born' inside a word as an itemized piece of lexical knowledge, defined once and for all, similar to atoms within a molecule. On the contrary, they will eventually emerge as possible patterns of linguistic use, as attested in text corpora.

Before turning to the main body of argument, in the next section we will outline the general architecture of the Italian Reference Corpus, which provides the necessary basis of our investigation.

2. The Italian Reference Corpus

The Italian Reference Corpus was created by the ILC and Mondadori (an Italian publishing house) in 1988. The work on the corpus is still in progress and is aimed at achieving a 'balanced' collection of journals, novels, handbooks, scientific texts, 'grey literature', etc., which will be a representative sample of Italian contemporary written language. A section on the spoken language is also planned.

At present the bulk of the corpus covers a period of six years, from 1985 to 1990, and contains about fifteen million words. It is divided into three subsets:

(1) The periodical subset (identification code: SI) which includes newspapers and magazines: total number of word-tokens 10,158,279.
(2) The book subset (identification code: LR) which includes novels, short stories, handbooks, scientific texts, etc.: total number of word-tokens 3,748,281.
(3) The technical subset (identification code: SR) which contains one-page technical reports on the projects of CNR Institutes: total number of word-tokens 1,142,998.

Concordances can be obtained using different strategies; the most common strategies are KWIC concordances which appear on printed outputs and/or can be displayed on the screen of a personal computer using the DBT system (Picchi, 1991) which interrogates the database in a fairly flexible and conversational way. KWIC lines are accompanied to the right by the reference, i.e. a set of codes subdivided into fields, which contains information about the source text.

Table 1a *Reference*

	A001-SI-PO-85-X				
Field	1	2	3	4	5
	Source				
		Subset			
			Topic		
				Year	
					Part of text

Table 1 shows an example of reference (note that field 1 codes vary depending on the subset code). In the following the codes used in each field of the reference are explained.

(b) *Newspaper headings*

Within the periodical subset (field 2 of the reference: code SI) we can easily extract newspaper headings and other related data:

Article Code	Pubblic. N.	Pubbl. Year	Newspaper Heading	Words N.
A001-A085	1	1985	La Repubblica	48089
...	...	...	...	...
A268-A272	1	1985	Casa Viva	8389
A273-A284	2420	1987	Grazia	15895
A346-A377	54	1986	Zero Uno	60315
A378-A390	7	1986	Cento Cose	17316
A423-A440	7	1985	Star Bene	32977
A441-A458	356	1987	Storia Illustrata	47317
...	...	...	...	...
3332-3349	405	1990	La Stampa	11343
3350-3392	33	1987	Il Mondo	34959
...	...	...	...	

(c) *Books*

Within the book subset (field 2 of the reference: code LR) we have a similar decoding list with bibliographical information:

Book Code*	Year	Author	Title	Words N.
M001	1989	Bobbio et al.	Corso di diritto	172341
M002	1989	Alberghina L.	Fondamenti di Biologia	23976
R002	1988	Bevilacqua A.	Una misteriosa felicità	95266
R004	1986	Arpino G.	Passo d'addio	33520
R005	1986	Forti M.	In Versilia e nel tempo	49393
...	...	...	...	...
T008	1988	Gage E.	Amore in terra	110557
...	...	...	...	...

* In the code field the alphabetical characters stand for M = Handbooks, R = books (in Italian), T = books translated into Italian.

(d) *Technical reports*

Similarly, references to source reports are provided in a decoding list for the technical subset (field 2 of the reference: code SR):

Report code	Year	Projects	Words N.
J001-K927	1987	Ordinary Projects: Prev.	552052
U001-V283	1988	Finalised Projects	84284
...	...	...	...
X001-Z999	1989	Ordinary Projects: Cons.	506662

(e) *Topics*

A similar decoding list is provided for field 3 (topic) as well. It displays in alphabetical order the acronym of the topic to which the text belongs and the subset in which the acronym can be found.

Achronym	Topic	Corpus Subset
AD	Furnishing	SI
AG	Agriculture	SI SR
AL	Nutrition	SI
AM	Environment	SI SR
AN	Anthropology	SR
AR	Arts	SI LR SR
...	...	...
PO	Politics	SI
PP	Psychology, Psychoanalysis	LR
PR	Celebrities	SI
PS	Sociology	
...	...	...
TU	Tourism	SI
UM	Humour	SI

(f) *Words per year*

For each subset statistical information is given about the distribution of words per year (see for example, the table for Newspaper Subset SI, where the number of processed articles per year appears):

Year	Words N.	Article N.	% (*)
1985	1778422	1801	17.51
1986	1721725	1413	16.95
1987	2474055	2460	24.35
1988	2530494	2223	24.91
1990	1653583	3239	16.28

* The percentage is the ratio number of words per each year/number of words of the whole subset.

(g) *Words per topic*

The distribution of words per topic and related percentage can be easily tracked down (see sample below from the table of topics present in the newspaper subset SI, sorted in decreasing order from the most to the least

frequent; the same data can be displayed also according to different criteria, e.g. in alphabetical order of the acronym):

Topic	Words N.	Article N.	%
PO	1323398	1817	13.03
EF	1210008	1573	11.91
SL	894133	1249	8.80
ME	781937	475	7.70
CS	670974	507	6.61
...			
FT	6251	4	0.06
MT	3236	5	0.03

The database can be interrogated and KWIC-lines obtained by a number of search keys according to different reference fields, interactively selected by a user.

3. Simple Statistic Techniques of Corpus Analysis

In this section we take a quick look at one type of data which can be easily extracted, using very simple statistical tools, from large quantities of textual data (for the Italian Reference Corpus, see Bindi *et al.*, 1989; Zampolli, 1990). The same types of techniques are described in Church and Hanks (1989) and in Calzolari and Bindi (1990).

The statistical tools which are used here are somewhat dated. However, whereas in the past these tools have been used mainly for stylistic and literary research, in the present paper they are used to study the language system. In our case, the extraction of particular types of information and their properties is couched, as stated before, in a general programme of research: namely, building a large computational lexicon.

The main type of data we extract are pairs of words which co-occur more frequently than expected in large quantities of texts. The measure of the strength of their association is based on the formula of mutual information given in the cited articles.

It is not necessary to insist on the importance of recording and taking into account collocational and co-occurrence data, especially for 'generation' purposes and translational correspondences across languages.

What we want to stress here is that, in order to make use of these word-pairs, it is necessary to make a linguistic classification of them within an appropriate theoretical framework, and that numerical results can be of help in defining a first 'typology' of the extracted co-occurrence data. The raw statistical data provide suggestions and hints which must await confirmation through linguistic analysis. Statistical methods must therefore be complemented by the input of linguists. Incidentally, a research programme aimed at developing such classification is in the making.

Another statistical tool which can be profitably used for classification of word combinations is the 'dispersion' index introduced in Calzolari and Bindi (1990), which measures, with respect to the keyword, the degree of fixity of the second word position in the selected window (see below for details). Its value ranges from 0, when the second word is always in the same position, to 1, when its occurrence is equally distributed among all the positions.

3.1 *Decreasing Mutual Information, Part of Speech Tagging and 'Dispersion'*

The first observation which can be made regarding mutual information values is that if we order the values in decreasing order we get very different types of co-occurrence data at different levels.

When the mutual information value is very high, the word-pairs share the following properties: (1) both words are of very low frequency, (2) in most cases they only appear together in the same context, not alone. This means that within the top range of values we find mostly proper names, foreign fixed expressions, compounds or co-occurrences belonging to specialized technical languages.

In the range of middle values we find more common, everyday words. In this case, if we want to be able to classify them, we have to introduce two other types of information, i.e. part-of-speech (POS) tagging and the value of 'dispersion' in the window. (Let f1, f2, . . ., fn be the number of occurrences of the second word at the first, second, . . ., *n*th position in the window. In the case at hand below f1 = g, f2 = h, f3 = i, f4 = 1. The dispersion index (D) is then equal to: $1 - V/\sqrt{(n^{-1})}$ where V is the ratio between standard deviation of f1, . . ., fn, and their mean (f). (For more detail, see Bortolini *et al.* (1971), pp. 23–31).) It is the combination of these three different types of information which enable us to go from raw data to more structured data.

From POS tagging we are able to distinguish the main types of combinations which are of interest at this level (Table 2).

Table 2

a	b	c	d	e	f	g	h	i	l	m	n	o	p
Noun	Noun		Cibi	Bevande	14	0	12	2	0	333	88	11.9	0.18
Noun	Noun	(PP)	Bisturi	Chirurgo	44	0	4	0	0	63	133	11.9	0.0
Noun	Adjective	(Mod)	Telefonata	Anonima	6	6	0	0	0	137	91	11.9	0.0
Adjective	Noun		Futuri	Assetti	4	4	0	0	0	168	50	11.9	0.0
Noun	Adjective	(Mod)	Principi	Attivi	29	29	0	0	0	394	155	11.9	0.0
Verb	Noun	(Obj)	Pagare	Bollette	6	1	2	2	1	481	26	11.9	0.81
Verb	Noun	(PP)	Quotare	Borsa	4	0	2	0	2	6	1390	11.9	0.42
Verb	Noun	(Obj)	Stringere	Alleanze	7	7	0	0	0	73	201	11.9	0.0

a, b, POS of first and second collocate; c, derived information on the second collocate; d, e, the two collocate words; f, total number of occurrences of the word-pair; g, h, i, l, number of occurrence of the second word in first, second, third, fourth position in the window with respect to the first word; m, n, total number of occurrences of the two words in the corpus (in any position); o, mutual information index; p, 'dispersion' index in the window.

We can then extract information from the concordances. For example, between *bisturi* and *chirurgo* there is always the preposition *del* and the noun *borsa* in *quotare ... borsa* is always preceded by *in*. We can therefore arrive at the conclusion that we have a noun or a verb followed by a PP (column c). Whereas in *pagare ... bollette* and *stringere ... alleanze*, the post-verbal position of the collocate noun is a clue to its being a direct object of the verb.

We can also see that the value of 'dispersion' (column p) informs us about the fixity of the expression. In the examples above we can contrast *pagare ... bollette* (where the noun is the object and can be found in any position of the window, according to whether there is or is not a determiner, and whether the noun is or is not preceded by a PP) to *telefonata ... anonima* which is a fixed expression (even though the modification relation is compositional and not idiomatic), or, better yet, to *stringere ... alleanze* where the object noun is always (though not necessarily, because the noun may be preceded by an adjective) found immediately after the verb.

Where lower values of mutual information are found, other types of pairs appear, i.e. syntactic/grammatical structures. Again it is the combination of the different types of information which give us hints for classification (Table 3). Using this method, we can combine lexical acquisition from statistical analysis and tagging to derive more structured collocational knowledge from the text.

3.2 *Right and Left Co-occurrences of a Keyword*

Another useful insight on the same data can be gained by simultaneously looking at the quantitative data related to both the left and right collocates of each word. Arrangement of the data in this manner is mainly of interest to the lexicographer. At the same time, we can also provide references to the source of our information. To be more specific, for our present purposes we have roughly divided our corpus into three main parts: newspapers/periodicals (P), books (B), and technical texts (T).

In Fig. 1 we show some collocates (chosen by automatically selecting only content words) of the word-form *corpo* (in English 'body, corpus, corps, staff' according to different contexts, leaving out translations of idiomatic phrases). The user is offered, on the screen or on paper, a synoptic view of the strongest collocations, which give us an immediate idea of the actual use of the word in its different meanings. For example, on the left, words such as *braccia, cellule, cura, esercizi, fitness*, etc., typically point to the 'body' meaning, while *comandante, comando, guardia*, typically go together with the 'corps' meaning, which is evidenced on the right-hand side by words such as *alpino, armato*, etc. Note that adjectives tend to show up mainly as right collocates and point to several hints as to the appropriate meaning of the first word.

In this type of display, values of mutual information are not only given by the corresponding figures, but are also visually represented by means of the relative geometrical distance shown on the screen (or on paper) between the collocate and the keyword. The user can interactively ask the system for such display, which can be usefully integrated into a lexicographic workstation.

The same kind of display can be obtained for those surrounding function words whose mutual information value is above a certain threshold. This gives the syntactic context of a word: see Fig. 2 for two forms of the verb *uscire* (English 'to go out'). Note that the types of prepositions on the right-hand side shed light on the most typical argument/modifier structure of the verb at stake (in the case at hand, given the intransitivity of the verb 'uscire', articles with no preceding preposition can be signals of 'subject inversion', a fairly frequent construction of Italian unaccusatives).

3.3. *Tagged or Not-tagged Corpus?*

As shown above, numerical results obtained by means of these raw methods need be integrated and refined with linguistic information. When the text is labelled with POS tags, we can, for example, (*a*) get the most typical categorial sequences for compounds, collocations, multi-word expression, etc., and (*b*) group together collocations concerning the same lemma (instead of working on inflected forms). Having said that, many observations must necessarily be made on specific word-forms. If we consider the word *contatto*, and take the verbs appearing on the left of the singular and plural forms, we notice that only three verbs out of twelve really apply to the lemma *contatto*. The other nine either co-occur with the singular or with the plural.

avere	contatt i
entrare	contatt o
essere	contatt o
evitare	contatt o
lavora	contatt o
mantenere	contatt o/i
mettere/rsi	contatt o
perdere	contatt o/i
stabilire	contatt o/i
tenere	contatt i
venendo/gono ...	contatt o

Table 3

a	b	d	e	f	g	h	i	l	m	n	o	p
Refl. pron.	Verb	Si	Avvelenò	5	4	0	0	1	72110	4	7.1	0.24
Poss. pron.	Noun	Sue	Affermazioni	8	7	1	0	0	3625	12	7.1	0.37
Pers. pron (obj)	Verb	Mi	Aiutato	20	0	17	3	0	8789	31	7.1	0.19
Verb	Prep.	Insediare	Alla	4	2	0	2	0	8	28838	7.1	0.42
Noun	Prep.	Conformità	Ai	5	2	0	2	1	23	12558	7.1	0.62

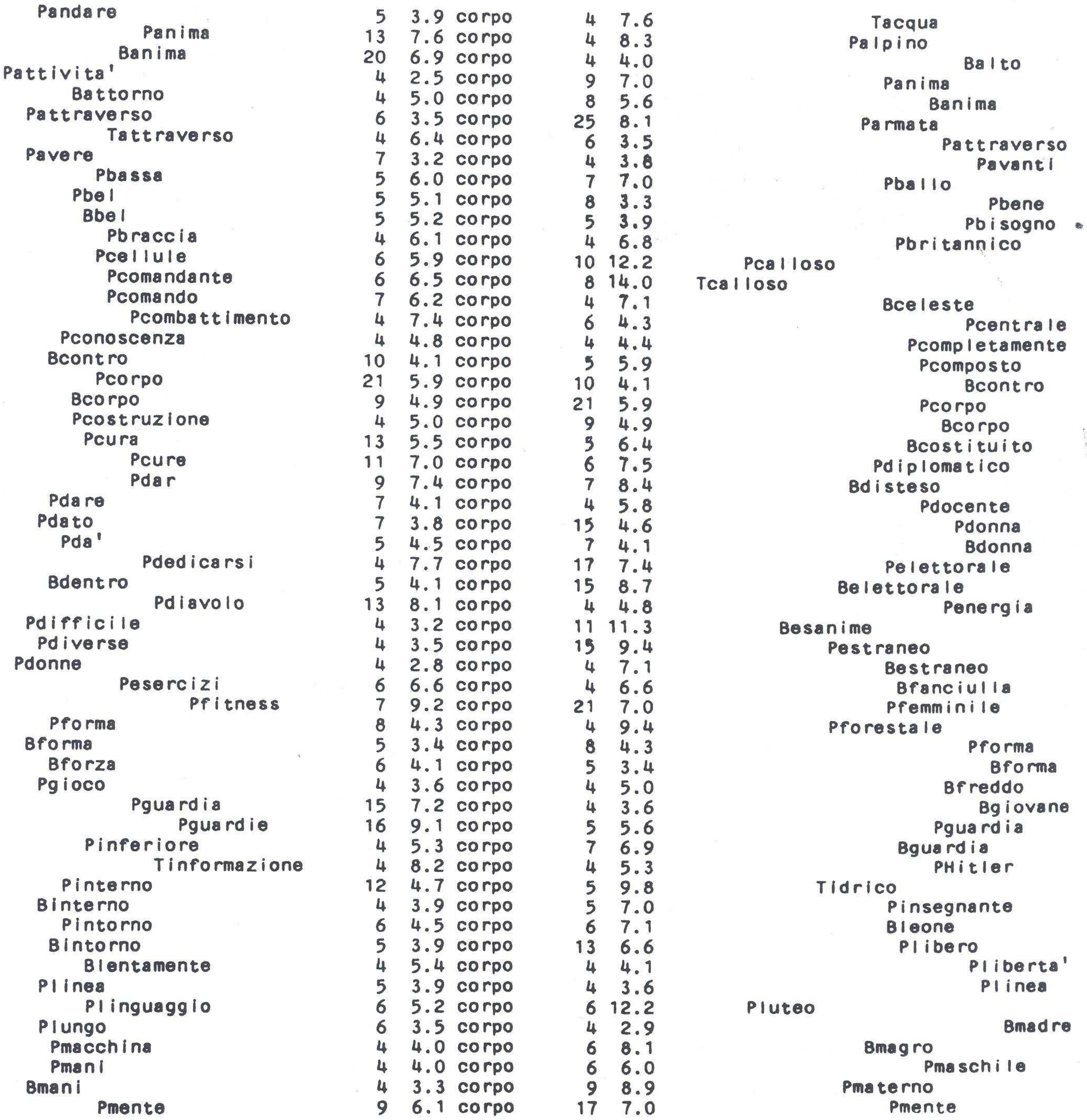

Fig. 1

These types of phenomena are very frequent, and must be recorded as such both in a printed dictionary (especially if it is a learners' or bilingual dictionary) and, even more crucially, in a computational lexicon.

We can continue the examination of this same lemma *contatto* by means of another simple quantitative tool which extracts all the tuples (from triples up to seven words together) appearing in the corpus more than three times. Looking, for example, at the quadruples with the lemma *contatto* in third position we see which of the above verbs (mostly verbs of movement used metaphorically) enter with this word into semi-fixed expressions (L indicates that we mean the lemma):

L tenere i contatti con
L venire a contatto con
L entrare in contatto con
L mettere in contatto con
L essere in contatto con

From these tuples we also get the information of which adjectives can modify *contatto* in the above semi-fixed expressions, as left modifiers:

in continuo contatto con
in costante contatto con
a diretto contatto con
a stretto contatto con

```
      Bdal                    4  4.6 uscirono      8  2.6          Ba
 Be                          16  2.7 uscirono      6  5.2       Ball'
    Bgli                      6  4.0 uscirono      4  4.7        Pcon
 Bi                           5  2.9 uscirono      4  3.0            Bda
       Bne                    4  5.1 uscirono      8  5.6     Bdal
          Bquando            11  6.0 uscirono     10  6.1  Bdalla
                                 uscirono      4  4.8        Bdue
                                 uscirono      4  2.8            Pe
                                 uscirono     18  2.8            Be
                                 uscirono      4  3.3           Pil
                                 uscirono      4  2.4             Ble
                                 uscirono      5  2.6            Bper
                                 uscirono      4  6.4  Btre
      Ballora                    7  4.8 usci'     4  4.5        Bcol
  Pche                          15  2.9 usci'     10  3.6          Pcon
Bda                             11  2.4 usci'     30  3.6          Bcon
    Bdal                        10  3.9 usci'      6  3.0            Pda
     Bdall'                      5  3.9 usci'     12  5.8   Pdal
   Bdalla                        6  3.3 usci'     33  5.6    Bdal
  Pdel                           9  3.0 usci'     10  4.9     Bdall'
       Pdopo                     4  4.8 usci'     10  5.9   Pdalla
    Bed                          5  3.7 usci'     45  6.3  Bdalla
Pil                             13  2.6 usci'      4  4.1       Bdalle
  Blei                           4  3.4 usci'      5  5.4    Bdietro
            Pne                 15  7.3 usci'      9  4.4      Bdopo
           Bne                  48  6.6 usci'     55  2.4          Be
     Pnel                       12  4.8 usci'     14  6.6  Bfuori
     Bnessuno                    4  4.8 usci'     12  2.5           Pil
   Bpoi                          6  3.6 usci'      5  2.2            Pl'
          Pquando               16  6.8 usci'      4  3.1         Blui
      Bquando                   24  5.1 usci'      4  2.9          Pma
      Bre                        4  4.8 usci'      7  4.0       Pnel
    Bse                         21  4.0 usci'      9  2.8          Bnel
                                    usci'      8  2.5           Pper
                                    usci'     23  2.8          Bper
                                    usci'      5  4.6      Bpoco
                                    usci'      4  3.0          Bpoi
                                    usci'      5  3.4         Bprima
                                    usci'      4  4.7      Bsubito
                                    usci'      4  3.3         Bsulla
                                    usci'      4  4.6   Psuo
                                    usci'      4  2.5        Bsuo
                                    usci'      4  4.6  Btutta
                                    usci'     11  2.9       Pun
                                    usci'     39  3.1      Bun
                                    usci'      7  3.6     Bun'
                                    usci'     25  2.9       Buna
```

Fig. 2

or also as right modifiers:

a contatto diretto con
in contatto diretto con

It is again the combination of different simple tools which helps in getting more and more detailed information.

This tuples extraction tool easily gives us information on idiomatic phrases (usually fixed or semi-fixed expressions) which must be recorded in NLP dictionaries (e.g. for translation purposes, given the non-compositionality of their meaning), but are not exhaustively listed anywhere. An example is:

L dirne	di tutti i colori
L diventare	di tutti i colori
L farne	di tutti i colori
L combinarne	di tutti i colori
L passarne	di tutti i colori
L vederne	di tutti i colori
L pensarne	di tutti i colori

L compierne	di tutti i colori
L inventarne	di tutti i colori
L scriverne	di tutti i colori
L conoscerne	di tutti i colori
L capitarne	di tutti i colori
L succederne	di tutti i colori

where we extract a large list of possible/actual variations of verbs co-occurring with the fixed expression. If we had to translate them into English where there is no literal corresponding of *di tutti i colori*, the translations must be quite different according to the accompanying verb

For example, in Collins we find:

diventare di tutti i colori	to turn scarlet
dirne di tutti i colori	to hurl insults at sb
farne di tutti i colori	to get up to all sort of hicks
passarne di tutti i colori	to go through all sort of problems

what about the others? and when two of them are combined as in: *'ne dice e ne fa di tutti i colori'*, or *'ne hanno dette e scritte di tutti i colori'*?

4. A Lexicographical Application

4.1 *Textual Corpus in Lexicography*

Sense distinction is a crucial task in lexicography: it may be very difficult to identify all the various meanings of a word and correctly describe them with significant exemplifications in a dictionary entry. Sense distinction of a word, when based on introspection and intuition, reflects subjective experience rather than the current usage of it in ordinary communicative situation.

Moreover, dictionaries, generally, tend to become an inventory of abstract statements about meanings: often, they encode all theoretical possibilities of a polysemous word without sufficient motivations for distinction in terms of collocations, syntactic structures, etc.; rarely actual frequency and current usage are recorded. Words appear torn off their context: the decontextualization produces a gap between the word itself and its meanings.

In computational lexicography, textual corpora are now known to be a precious tool and a significant source of knowledge: they can help to investigate more correctly 'the world of a word', on the basis of textual evidence (how a word is pragmatically used in texts).

The context in which an item appears offers, moreover, what is 'typically' used rather than 'potentially'; in this way, the context permits to discover:

- Contextual restrictions 'spotting' the current meaning among all the possible.
- Significant regularities linked to a sense distinction.
- Objective criteria in distinction of subtle and salient senses (see Cotoneschi and Monachini, 1991).

A methodology based on these observations has been adopted in a project of multilingual translation equivalences, aimed at tackling problems of word-meaning distinction across languages: by concentrating on all recurrent environmental patterns, particularly those which appear linked to different meanings, we can formalize them as 'restriction' rules and claim that automatic disambiguation routines can be successfully devised (see Calzolari, Cotoneschi and Monachini, 1990; Sinclair *et al.*, 1990; Monachini and Calzolari, 1991).

Often concordances offer too roughly arranged linguistic facts with the result that textual evidence is not 'so evident'! Concordance analysis and disambiguation criteria design seem to be extremely laborious and to cause omission. A significant phenomenon can be spread over a great number of occurrences and it cannot be always easy to set up significant rules for distinguishing a particular sense rather than another.

Statistics, by processing textual data with appropriate criteria, helps to select relevant linguistic phenomena. Naturally it alone does not solve all the problems but, by grouping and ranking data in order of importance, paves the way to human intervention: the attention of lexicographers, for example, is correctly focused on salient linguistic facts, which are worth representing in a dictionary entry, while more remote possibilities can be easily omitted (see Church *et al.*, 1990).

Mutual information index, by giving the measure of strength of association between words is also useful to identify, in the context of a word, sets of candidates that help to disambiguate among various senses and to translate a given pattern into a formal rule, in order to obtain a more objective disambiguation.

We give below only some general applicative examples of the analysis carried out both on the Italian Reference Corpus and on some Italian dictionaries.

4.2 *Convergence of Statistic Techniques and Rule-based approach in Sense Disambiguation*

The idea we want to propose in this section is the application in lexicography of both statistical processing of textual corpora and sense distinction in terms of formal rules. We give here some examples of the most frequent types of ambiguity and problems of sense selection, which lexicographers are concerned about.

CIGLIO

The analysis of the word-pairs obtained by processing the corpus with mutual information formula suggests typical collocations linked to the distinct senses of the noun *ciglio* which in Italian is ambiguous between (1) 'edge' and (2) 'eyelash'.

(1) The first sense is easily disambiguated by its co-occurring with nouns belonging to the same semantical field:

ciglio burrone	13.8
ciglio strada	9.4

(2) In the second sense *ciglio* is attested in the corpus (and presumably used) only in the fixed phrase represented by the word-pairs:

battere ciglio	14.2
senza ciglio	8.8

The whole locution in the couple *senza ... ciglio* is *senza battere ciglio*, which pictures a reaction of impertubability in front of something unexpected. In the newspaper subset this locution has a higher strength of association in the more colloquial truncated variant:

batter ciglio	17.2

The sense of 'eyelash', in Italian, is attested specifically in the plural form, a particular plural which comes from an ancient Latin dual (as happens for almost all the terminology of double parts of body) with the desinence *-a*:

lunghe ciglia	10.2
ciglia finte	14.7

In the plural form *ciglia* is polysemous between 'eyelashes' and the technical 'cilia' ('filers'): this last sense is represented by the strong pair:

ciglia flagelli	14.7

By looking into concordances, we find contexts of this item in the sense of 'cellular appendixes' labelled with the code of the corpus subset devoted to technical and scientific sublanguage. Moreover, the reference system, which contains (together with other information on the source of the text) the topic of the text where the item occurs, offers some additional information on the sense distinction linked to specific lexical field: the second sense is more frequently used in biological and medical field.

The consultation of unprocessed concordances and their integration with processed data is often very useful.

All these facts have to be represented and described in the dictionary: typical usage of plural instead of singular, semantic constraints in the choice of two different plurals (*cigli* for the plural of *ciglio* = 'edge'; *ciglia* for the plural of 'eyelash'), technical usages, particular fixed phrases (the comprehension of which may be difficult), are important in monolingual dictionaries but especially crucial in bilingual and learners' dictionaries. This type of information have to be introduced NLP lexicons.

AIUTO

It is even more interesting to investigate and try to solve with the two approaches polysemy in terms of subtle sense selection; in such cases it is more difficult, but still possible, to discover recurring patterns and to translate them into rules.

For the noun *aiuto* which means 'help', in the moral sense, and 'subsidy', in the material one, we want to detect if the context suggests some patterns characterizing the two distinct meanings:

(1) The plural form *aiuti* typically co-occurs with adjectives pointing to the material sense (on the right):

aiuti	economici	8.4
aiuti	finanziari	9.0
aiuti	militari	8.3

(2) The singular form co-occurs, more frequently, with verbs, such as 'chiedere' (on the left):

chiedere	aiuto	8.8
chiedono	aiuto	7.8
chiede	aiuto	7.5
chiedeva	aiuto	8.4
richiesta	aiuto	6.1

The morphological criterium singular/plural can be formalized in rules which distinguish the two senses:

chiedere+aiuto (singular) = (moral) 'aid, help'
aiuti (plural)+adjective = (material) 'subsidy'.

This intuition of sense distinction linked to the opposition singular versus plural has to be integrated into dictionaries with the specific collocations.

PIATTO

We want to detect now the problem of omography between noun and adjective, such as in Italian *piatto*: (1) the adjective 'flat' and (2) the noun 'dish': the noun is polysemous in the sense of (*a*) 'container of food' and (*b*) in the metonymic one of 'food'.

(1) The adjective is accompanied in left co-occurrence by nouns which can have typically the characteristic of being 'flat':

fronte	piatto	5.2
schermo	piatto	8.2
tetto	piatto	8.7

In this sense is also attested the fixed expression which refers to a physical fault: *piede piatto* 8.4, *piedi piatti* 7.4.

(2) The senses of the noun are disambiguated in the word-pairs by co-occurrence with typical verbs (to the left) or with other characteristic kitchen utensils (to the right):

(*a*)	lavare	piatti	11.5
		piatti bicchieri	12.0
		piatti posate	12.1

In this sense the expression *piatto della bilancia* (11.0) has a high mutual information value.

The sense (*b*) is represented by the co-occurrence with (i) verbs and (ii) adjectives typically related to food, and (iii) nouns synonyms and hyponyms of 'meal' or 'food'.

(*b*)	(i)	gustare	piatti	11.2
		mangiare	piatti	7.7
	(ii)	primi	piatti	7.3
			piatti gustosi	12.9
			piatti tipici	10.2
			piatto caldo	7.6
			piatto forte	7.7
	(iii)	cibi	piatti	8.3
			piatto minestra	10.8
			piatto pastasciutta	12.1

RIPARTIRE

Semantic ambiguity of verbs is often solved in terms of the opposition content words/function words. The verb *ripartire* means 'to divide' but it is also the iterative form of the verb 'to leave', i.e. 'to leave again'.

A straightforward rule for sense distinction relies on a rough and ready syntactic criterium: namely, the opposition transitivity versus intransitivity. Couples obtained with the mutual information index are crucial indicators of the two meanings: typical transitive and intransitive uses are immediately identified by looking at the syntactic context following the keyword:

'dividere'	ripartire	costi	9.4
	ripartita	tra	7.1
'partire di nuovo'	riparte	da	4.1
	ripartito	per	4.7

In this respect, the index of dispersion, which gives us information about the relative position of the second of the two words in the selected window, helps a lot to assign an appropriate syntactic function to the former,

and thereby to correctly resolve ambiguity which bears on function words and synctactic criteria.

4.3. *Some General Remarks*

Statistical analysis of textual corpora gives lexicographers the possibility of knowing what 'to look for' in concordances: textual evidence is turned into statistically significant textual evidence. A contextual rule-based approach helps to disambiguate meanings of words on a more objective basis. The aim of this experiment is to prove the efficacy of the combination of the two above proposed approaches in order to:

- Obtain satisfactory and correct sense distinctions.
- Integrate existing dictionary entries with the knowledge coming from current usage.
- Create the basis for newly structured printed dictionaries.

In particular, the knowledge extracted through the analysis of large bodies of current texts has to be integrated into computational lexica for NLP purposes. In the long run, we plan to deal with problems of lexically conditioned choices among modifiers belonging to the same semantic field, which is a well-known 'crux' especially in bilingual lexicography. Furthermore, we intend to integrate the techniques described above with multivariate statistic techniques working on groups of words (see the following section).

5. Multivariate Statistic Approaches

5.1 *Nature of the Problem*

In this section we move towards a more comprehensive approach to word-meaning, and, hopefully, to word-semantic theorizing itself.

Lexical restrictions over the range of permitted arguments, idiosyncratic collocations, preferences and co-occurrence constraints of various sorts can hardly be viewed as peripheral aspects of word-meaning. In our view, they just mirror, arguably in the most extreme fashion, the inherent complexity of the more general problem of assigning a sense to a given word. Our model will hopefully help us to answer the following basic question: how can one explain the feeling that a general word sense or word concept appears to have '. . . an integrity or wholeness, which current representational schemes are unable to capture directly' (Kelly and Stone, 1975, p. 74), in the face of the comparatively motley bunch of its lexical uses in actual contexts?

The prevailing myth of being confronted, in ordinary circumstances, with clear-cut characterizations of word meaning (allegedly the solid core of each given word sense), is soon dispelled, when the evidence provided by ordinary dictionaries is brought under closer scrutiny: for each pair of sense characterizations listed therein for a given entry, one can easily find usages falling roughly continuously between them (Kelly and Stone, 1975). Such apparently seamless, elastic fields of word senses, stake out fuzzy sets of meaning which appear to fade into each other, overlap, and collide along exceedingly finely grained borderlines. To give an example, the difficulty of outlining the lexical boundaries between a pair of quasi-synonyms like 'strong' and 'powerful' (Halliday, 1975; Church, 1989–90) points directly to the problem of the internal organization of word senses in the conceptual system of a language user. The problem is that the difference in syntagmatic distribution between *strong* and *powerful*, i.e. the set of words they do keep company with, is apparently too idiosyncratic to be captured by the logic grid of a deductive, paradigmatic approach. As Halliday puts it, '. . . the paradigmatic relation of *strong* to *powerful* is not a constant but depends on the syntagmatic relation into which each enters'. Word senses behave differently in different contexts.

Church (1989–90) has carefully investigated the phenomenon of *lexical collocations* from this viewpoint in the framework of a statistic approach, to arrive at the same sort of conclusions Halliday drew on a more theoretical basis. Put in a nutshell, the tricky thing about co-occurrence restrictions between pairs of words and similar lexically-driven, context-sensitive constraints, is that we cannot so easily make absolute statements about them, when it comes to negative evidence. More concretely, this means that, given an anomalous pairing like 'powerful support' as opposed to 'strong support', one cannot prove the former to be linguistically anomalous by pointing to the fact that it is, statistically, highly unlikely to occur in normal contexts. This because, in many cases, the best result one gets is that the anomalous pairing at stake turns out to be statistically unlikely, but not that unlikely for us to rule out, as it were, its 'right' to ever turn up again. This outcome is arguably due to several factors which we will not go into here. Church's way out is to resort to the *t*-score test. The essential point here is that the unlikelihood of a sequence like *powerful support* is not to be gauged in absolute terms, but against the background, as it were, represented by the set of possible lexical alternatives we have at hand in our mental dictionary. In other words, the stress must be laid on the fact that we normally use *strong* instead of *powerful* when *support* follows. This is exactly what stated by Halliday, which can be complemented by the following point: the oddity of the distributional properties of *powerful* and *strong* is in focus only when set off against the 'stage set' of their (quasi)-synonymy, which defines the *standard* modality of their use. All of which leads us to the following points:

(1) It seems impossible to fix, a priori, the number of meaning dimensions along which a given word 'stretches' its sense(s).
(2) It seems exceedingly difficult to define the link between words in the mental lexicon independently from their distributional behaviour in contexts.
(3) An answer to the problem of lexical relatedness is expected to be as comprehensive as possible: what holds for a couple of quasi-synonyms, must hold for the whole system of their co-ordinates/synonyms as well; it is unnatural to seek for a piecemeal explanation, when the goal is to formulate general principles.

Having brought these points home, two further questions naturally arise: how can we 'mould' such apparently cahotic 'entropy' of word senses into a more perspicuous, self-consistent representation? How do we compute any further possible extension of their use?

5.2 *The model*

The first step is that we turn a 'conceptual' problem into a 'geometrical' problem. So far we have been talking about meaning dimensions in a figurative way; now we are taking them literally. In our model, an abstract *n*-dimensional state space is represented. Therein a metric defines the relationship between two possible positions in it, in terms of a geometric distance. Every word type takes a particular position. Given a bundle of word types whose conceptual links we want to investigate, for each pair of them we can state: (1) whether there is a third unit which falls in between them (or even more than one); (2) which of them is closer to a further position taken as a point of reference; (3) which of them is ranked higher in relation to each dimension; (4) along which dimensions(s) they are opposed to each other.

Intuitively, we build up a chart, by ranking each item with respect to each dimension. When we want to check the similarity between a pair of them, we just scan through the chart, calculate the geometrical distance between the positions taken by the two items, and give the answer. Our chart differs from any semantic hierarchy in virtue of its multidimensionality: similarities and dissimilarities do not need to be stated once and for all, and independently from a particular perspective. Items which come very close to each other along one axis, can turn out to be strewn far away from each other, along another axis.

Having defined our representational model, we now move on to characterize the computational problem. Within a traditional, hierarchical model, the need to account for the multifarious set of possible relevant oppositions among a selected group of items, led to construing a different hierarchy for each different perspective. This is no longer needed within this framework. Somewhat paraphrasing Paul Churchland (1986), a specific state of affairs within an appropriate group of word senses, i.e. the conceptual picture of their relationships, is just a particular, purpose-orientated projection of an underlying multidimensional data structure. It is as though one cut a 'slice' through a massive, highly structured amount of information: each slice is a snapshot of the current lexical space, taken from a particular angle. 'The collective coherence of such sample slices is a simple consequence of the manner in which the global information is stored at the deeper level' (Churchland, 1986, p. 364). This has an important consequence: adding one further dimension to the overall state space affects the configuration as a whole; therefore it is bound to alter, at the same time, the way sample slices will look like. In other words, every single further pairing instance that we add to our data structure, does not make the latter grow up by accretion; rather, it actually restructures the entire data organization, by recalculating distances between points. Thus, our model accounts for the way fresh contributions refine the original competence model. Put another way, paradigmatic structures are not a priori representations; rather, they are built up from syntagmatic relations, and are open, as such, to permanent revision. Finally, the problem of extending the use of a word to other contexts which are similar to those already in use, can be looked at as the search for the shortest path from that word to a new lexical item; this does not need to be done in one fell swoop: intermediate points can be used as 'pit-stops' before a suitable goal is got at (on which more below).

Algebraically, cutting a sice through a multidimensional data structure is like projecting *n*-dimensional positions on to a surface. The slant of such surface is the angle the blade of an imaginary knife forms with the surface of our multidimensional pie. Formally, the problem amounts to a multidimensional scaling from the original *n*-dimensional system of our data structure down to a more manageable, generally two-dimension, 'pocket' representation.

5.2.1 *Input and Output.*

First, we are confronted with sequences (strings) of word tokens in contexts as they show up in text corpora. Our final objective is to build up a semantic hyper-space where semantic dissimilarities between word-types are represented as proportional geometric (Eucledian) distances between points. Each point stands for a relevant word-type. Proximity between points in this hyper-space 'represents' conceptual affinity; remoteness, conceptual dissimilarity.

How do we map 'raw' strings on to paradigmatic representations? We do this by means of scaling algorithms. Intuitively, they elicit stereotypical lexical information from the open-ended variety of lexical usages in contexts. Rieger (1991) calls these algorithms abstracting functions, since they actually extract generalizations from unrelated data. This will become clearer as we progress in presenting our sample application.

5.2.2 *The data.*

For the example we present here, we have focused our attention on a closed, intuitively consistent, set of Italian words which can be said to generally point to the semantic area of smallness, in a more or less strict dimensional sense. Our set was initially chosen in the light of an explorative survey which seemed to offer interesting results.

The words under scrutiny are the following: *piccolo, corto, breve, ristretto, esiguo, scarso, ridotto.*

In some cases, some of them appear to be used interchangeably with one another. What we wanted to investigate is whether such dissimilarity of contextual behaviour can be profitably exploited to characterize their semantic relationships. Eventually, the overall, multidimensional structure emerging from this type of information can be regarded as a semantic hyper-space of a certain subset of the dictionary: close synonyms cluster together in the geometrical plot; less similar words tend to take a more peripheral position. More-

over, the core of our derived configuration is expected to be taken by those words which prototypically characterize the conceptual space in question, in analogy with Rosch's view that people do not categorize common objects by arranging them on an equal footing, but by picking up an ideal exemplar, or 'prototype', and by matching its features on to other items' (Rosch, 1975; Kelly, Bock, and Keil, 1986).

5.2.3 *Units and Parameters*. From now on, we will technically refer to the quasi-synonyms listed above as units or individuals. Units are what we want to give a picture of. Since our method is relational by definition, units must be characterized with respect to something else (e.g. the axes of a hypothetical hyper-space). Since the techniques we illustrate here are multivariate, each unit will be defined by means of more than one parameter. Oversimplifying a bit, parameters are, following Church, the unambiguous words each unit does keep company with in ordinary contexts (e.g. every word *piccolo* keeps company with: *bambino, quantita', dimensione* and the like): parameters provide us with the syntagmatic information we need, to build up our semantic model. Less formally, these are what we will refer to, in the following, as 'word mates'. By selecting an appropriate set of parameters, we set up the relevant criteria as how to abstract away from the linear dimension of texts, to get to the more abstract, multidimensional level of typical paradigmatic relations. Parameters are chosen by calculating the mutual information index between each unit and each word the former occurs with, and by picking out the most correlated 'word mates' to each unit.

5.2.4 *First Stage*. Intuitively, two units (word-types) will be the closer to each other in the semantic hyper-space, the more similar is the range of relevant words they usually occur with ('word mates').

At a first level of abstraction, we will therefore extract from texts a matrix $n \times m$, where n is the number of units, and m the number of parameters. For each row i, there is a vector u_i representing the ith unit of our set. Each slot u_{ij} for $1 \leqslant j \leqslant m$ contains a co-occurrence value for both the ith unit and the jth parameter, a value representing pairwise relatedness between word-types.

To a first, fairly rough approximation, such a value can be looked at as the co-ordinate which expresses the ith unit with respect to the jth axis. Clearly we can possibly have many different ways to calculate such coordinate. Our choice at this level of abstraction affects following steps. In most cases u_{ij} expresses (an estimate of) the joint probability $P(W_j,p_j)$, where W_i and p_j are respectively the ith word-sense (unit) represented and the jth parameter, with $1 \leqslant i \leqslant n$, and $1 \leqslant j \leqslant m$. As we hinted at before, in our sample application u_{ij} is the mutal information value.

In the present context, the distinction between units (or individuals) and parameters is fairly artificial, since both are represented by word-types, and is justified only on the basis of the different role they perform. This contrasts with ordinary applications of multivariate methods where individuals are often characterized by parameters of a different nature (e.g. a sample of people, described according to their weight, height, limb length, etc.). This fact makes those multivariate techniques which represent contextually both units and parameters in a hyper-space, even more interesting for our purposes. We will turn back to this point later on.

5.2.5 *Second Stage*. One can view the $n \times m$ matrix as the fuzzy hyper-space of paradigmatic relations among the relevant units we are aiming at. Axes of the hyper-space are represented by parameters, which in turn document dissimilarities of usage among units (note that, in some cases, u_{ij} values may not behave as a metric, in a strict mathematical sense; they may therefore need to be manipulated accordingly). The trouble is that, ordinarily, we need several parameters to give an accurate characterization of each unit, which means that we are usually confronted with a hyper-space staked out by m axes, with m fairly large. Put another way, our $n \times m$ matrix gives us too much information to be of any use: it has to be shrunk down to a more manageable dimension.

In statistics, a number of scaling algorithms have been designed to achieve this purpose. There are as many different multivariate techniques as reducing algorithms. In the following we will informally hint at some of them.

Multidimensional scaling (MDS). The acronym MDS stands for multidimensional scaling. It encompasses a whole bunch of stepwise, iterative computational techniques which achieve the common purpose of shrinking a multidimensional data structure down to a more manageable representation. Usually, but not always, the outcome is two-dimensional. Multidimensional scaling is fairly young if compared with 'numerical' strategies (like factor analysis) which go as back as to the beginning of this century, but has been more and more used over the last few years, especially in processing data elicited through perceptive experiments. The objective is to scale down the dimensionality of an original set of data. Since the algorithm is iterative, the MDS will try to get at the smallest possible configuration which meets the constraints over data, where 'smallest' refers to the number of axes.

Intuitively, constraints are defined by how similarities among units are mirrored by geometrical distances between points. The larger the dimensionality of our space is, the more easily we can make similarities fit in with geometric distances (since we have more degrees of freedom for points to be rearranged). As we go on dropping dimensions, fitting gets harder and harder. For each solution, a value of stress will tell us how well distances match similarities. The method is very attractive for its generality. Unlike other methods (see below), it does not seem to impose preliminary hypotheses on the data. For our example, we have chosen a particular version of MDS.

Factor analysis. Factor analysis looks for the commonalities conveyed by data. Units (word-types) are assumed to share a hidden, underlying structure, which factor analysis is expected to bring back to surface. Data are processed so that differences among indi-

viduals are given a low score, and taken as peripheral aspects which do not affect the core of similarities. Commonalities are reinforced (mathematically, covariance between units is maximized), and the output ends up giving a picture of the strongest associations among the word senses at stake; differences are not blotted out, just considerably 'muffled'. To be more concrete, factor analysis can be regarded as the abstracting process by means of which details are discarded for the sake of generalization: it is this process which allows us to consider, e.g. *piccolo, breve, corto, scarso*, etc., as a set of quasi-synonyms, in spite of their differences in ordinary communicative situations. Idiomatic expressions are therefore not given prominence: a word-form like *farla* which is strongly correlated to *breve* within the phrase *per farla breve* (English 'to cut a long story short') would take a back seat. Such an abstraction reflects the intuitive appeal of the idea that the 'core' of meaning of *breve* is what the latter shares with *corto, scarso*, and the rest, rather than what is typical of *breve* and not of, for example, *corto*.

Analysis of correspondences. This is a fairly straightforward and efficient algorithm which is ideally applied to contingency tables (i.e. two-entry tables where each cell indicates the number of times the 'row-word' and the 'column-word' occur together).

The most interesting feature of such technique is the output format. On the two-dimensional space which is outputted, both units and parameters are projected at the same time, in such a way we can simultaneously analyse both the most relevant dissimilarities among units and how parameters contribute to such configuration. Actually parameters are shown as points distributed around their own unit, as satellites orbiting their own planet.

Such a projection pictures every word-type as a bundle of different 'word mates' (parameters, each of which linked to it by a different association strength), rather than a simple, discrete, isolated entity of some kind. The core of this word-type is given by the distributed pattern of those 'word mates' which are most likely to occur together with it in ordinary contexts; its kinship with other word-types is represented by the set of 'satellites' they share in common, i.e. those points which are shown, in the semantic space, halfway between the two.

Discriminant analysis. In Discriminant Analysis, quantitative data (e.g. frequencies, indexes of relatedness and the like) and qualitative differences discriminating them, are simultaneously accounted for. Given some units (e.g. quasi-synonyms as *piccolo* and *scarso, breve*, etc.) and parameters (e.g. co-occurring words like *bambino, capelli, dimensione*, and the like), we can ask ourselves how many different senses can be assigned to units. In our opinion, a question like that makes sense if and only if some, even fairly gross, conceptual standpoints are given in the first place. To give an example, once we know that semantic feature ±abstract is a relevant standpoint, then we can just pick up different senses according to their discriminating 'power' with respect to the opposition ±abstract. Accordingly, we can, for example, assign two senses to the word *stretto*: one, when *stretto* is followed by a +abstract noun-head (as in *senso stretto*, *stretta correlazione* etc.), the other when it is followed by a −abstract one (as in *fascia stretta*, *strada stretta*, etc.).

Clearly, this is rather crude (what about *parente stretto*?). But we can easily multiply lexical/semantic constraints on the set of potential noun-heads of *stretto*, ending up with a by far more finely grained classification of related word senses.

What is important to stress here is that, given this standpoint, different senses are not concepts contained by a word; rather they are perspectives over lexical restrictions of different degree of strength. In our model they are represented as the axes which best group together those words bearing the same semantic feature value(s), by keeping apart those bearing incompatible feature values. In a nutshell, word senses are ways of structuring our lexical competence, conceptual grids superimposed over an otherwise continuous blend of lexical links. This approach does not exclude feature-based information altogether; it just puts it in a different context.

This is exactly what discriminant analysis is for: it is achieved by minimizing variance within groups of omogeneous words (i.e. characterized by the same semantic features), and maximizing variance among groups.

Note that we are not postulating only one way of discriminating word senses, once and for all: clearly we get different outputs, when different standpoints are made interact with the whole picture.

5.2.6 *Configurations and comments*. As already pointed out in the paragraph above, units and parameters play symmetric roles, since the relation between word-types and 'word mates' is expressed in terms of co-occurrence frequency, which is a symmetric relation.

This means that the configuration of word-types as derived through a certain set of parameters, can be easily converted into a configuration of parameters, by given units. Usually the two complementary configurations are represented one at a time.

The configuration we illustrate here has been derived by applying a variant of standard MDS to our data. Both structures, units by parameters, and parameters by units, are projected on to the same plane. The interpretation of it is fairly straightforward. First, the distribution of points representing the selected word-types (*piccolo, scarso, ridotto*, etc.) pictures the configuration of the semantic space of 'smallness'. Secondly, each word-type can be intuitively viewed as the centre of gravity of the bunch of 'word mates' which 'swarm' around it. The closer a 'word mate' is to its word-type, the higher the mutual information index between the two.

To sum up, the structure we obtain this way is the result of a threefold set of constraints: (1) given a pair of word-types, if both are usually accompanied by the same 'word mate' (or even more than one), they are represented close to each other; (2) given a pair of 'word mates', if they do accompany the same word-type (or even more than one) in contexts, they show up close to each other in the output configuration; (3) given a pair of one word-type and one 'word mate', they stick

together in the output plot, if their index of mutual information is high.

Having defined this set of constraints, it is easy to understand that, when they are taken into account for each pair of units and parameters at stake simultaneously, points cannot be possibly scattered at random, but have to be harnessed into a pretty well-defined, coherent structure. Which means that each position is open to interpretation. This is indeed an attractive feature of this representation.

Figure 3 shows the structure of the semantic space of 'smallness', as shaped up by the data in our possession. It looks like a letter 'y' standing on its head. This is a fairly unusual shape for standard MDS outputs (most of them presenting some kind of horseshoe-like pattern), meaning that word-types are differentiated fairly well along the dimensions we chose. This is not surprising, since, as already said, we picked up the most correlated 'word mates' for each word-type, by the mutual information index. What is interesting to note in the first place is the emerging of a natural 'core', which groups most of the units at stake, plus three, distinct sort of 'feelers', sticking out of it: one, bearing the heading *corto*, right at the top of the diagram; the other two,

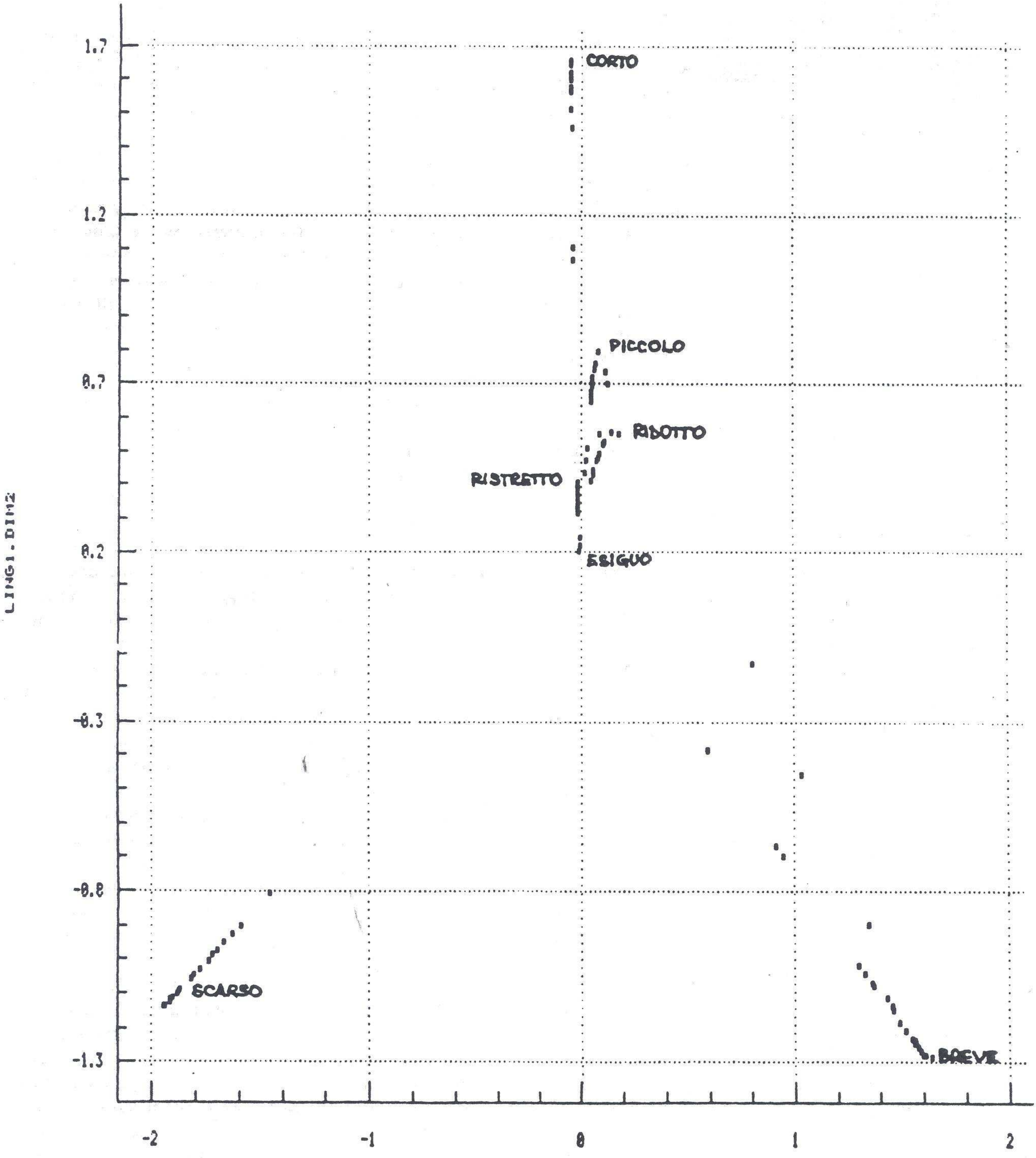

Fig. 3

scarso and *breve*, respectively, at the bottom left and right corners of the figure. We already anticipated the most straightforward interpretation of the 'core': it tends to contain, for a carefully chosen bunch of words, those among them which prototypically represent the semantic area in question. Indeed, *corto* (short), *breve* (brief), and *scarso* (scanty) are intuitively more peripheral to 'smallness' than 'small' itself. More interestingly, this first-glance interpretation is corroborated by a careful, closer scrutiny of each represented sub-area.

Figure 4 is a blow-up of the central, fairly crowded area of the picture. The blow-up allows us to focus on the semantic space staked out by four word types: namely, *piccolo, ristretto, ridotto, esiguo*. Let us begin with *piccolo*. A good way to interpret this type of geometrical configuration is to be on the lookout for keywords or thematic words, like quantity, length, number, and the like. They are obviously strong candidates *qua* prospective semantic axes, pivotal concepts which shed their light on surrounding words. Among the several neighbours of *piccolo*, we find a number of interesting keywords: *quantita', numero, set, dimensione*, and *formato*. They jointly point to a broad con-

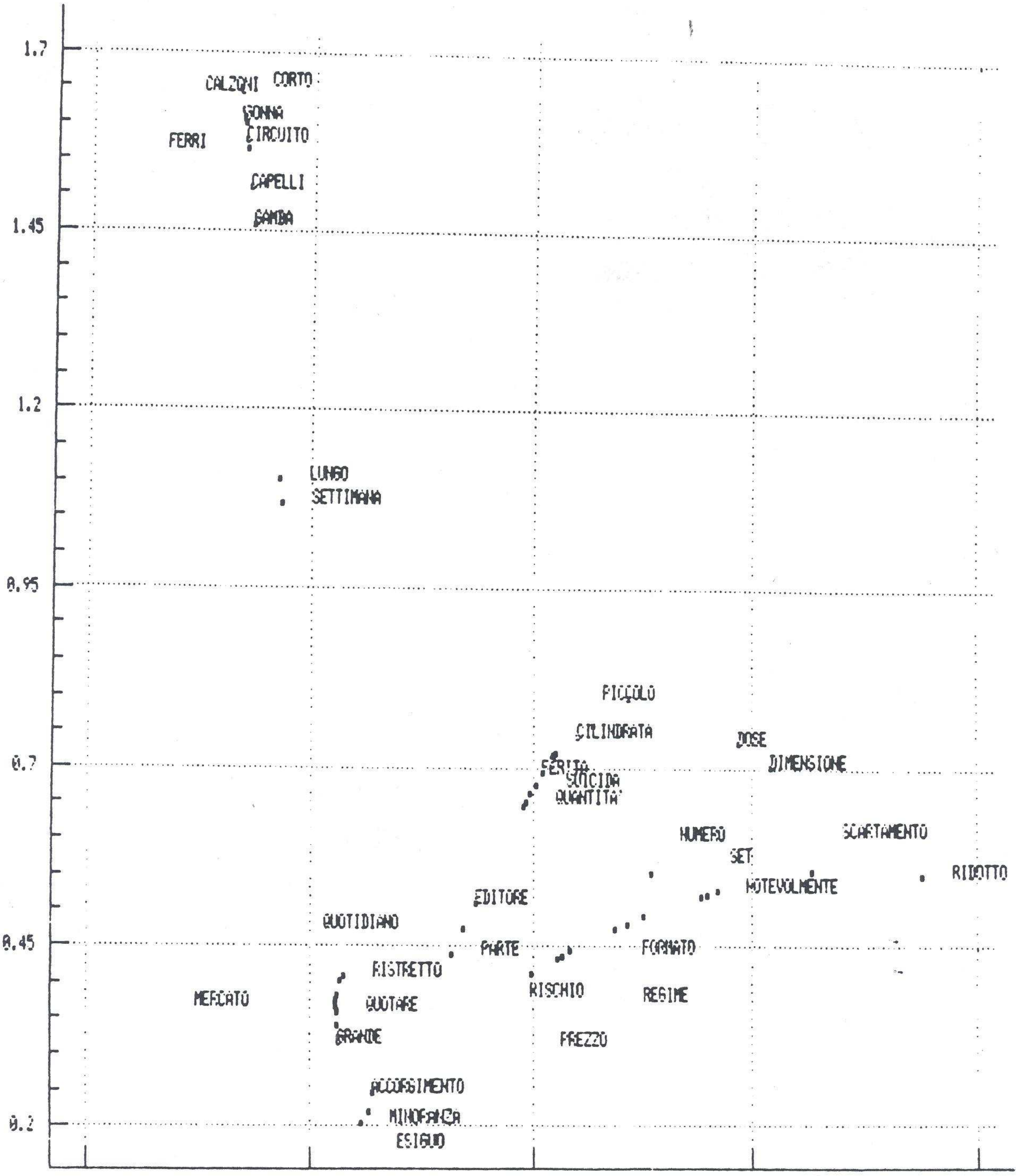

Fig. 4

ceptual area, where quantity, countability, and dimensionality play a crucial role. This is certainly a rather trivial remark if taken out of context. It looks much less so, however, if one considers (*a*) that this nucleus has emerged from a whole lot of unconstrained textual evidence, (*b*) data preprocessing is limited to eliciting fairly simple co-occurrence indexes, (*c*) the nucleus itself is opposed to three other nuclei, each of which pretty well characterized. It is interesting to note the presence of the adjective *grande* in this area. The presence of such co-ordinates/antonyms is always instructive since it shows what other words are frequently picked up and coordinated in the same contexts. Note, for another example, *lungo* halfway between *piccolo* and *corto*. Clearly, its presence signals that, heading north-west, as it were, we are leaving the realm of 'quantity' to enter that of 'length'. Namely, the second nucleus that one finds along that direction is headed by *corto* (short). What is more, all 'word mates' cluttering that area refer to concrete objects (*gamba, calzoni, calzoncini, gonna, capelli*, etc.).

Diametrically opposed to *lungo*, with respect to the central cluster, one finds *medio*. Somewhat surprisingly, *medio* is not directly linked to *corto*/*lungo*; rather it

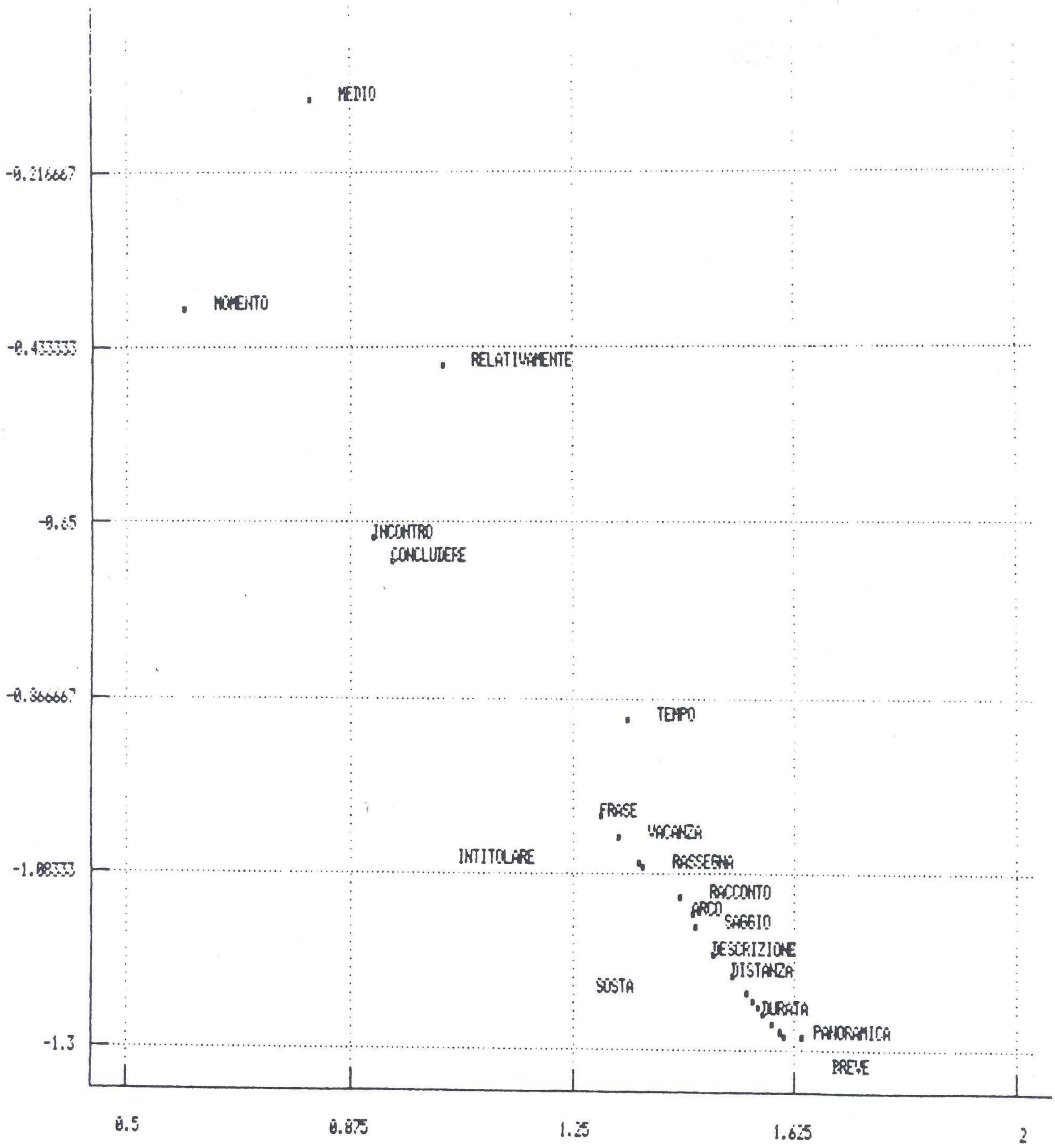

Fig. 5

takes an area halfway between *piccolo* and *breve*. This is due to the fact that *medio qua* parameter ('word mate') is more frequently associated to either *piccolo* or *breve* than to *corto*. Arguably, *medio* would be represented quite differently if another set of coordinates were taken into account. A more interesting case of 'unexpected' result is *settimana* (week). Its position is close to *lungo* (Fig. 5). Now Fig. 5 shows a blow-up of the cluster in the right bottom corner of Fig. 3. The word-type in question is *breve*; *breve* is commonly associated to, among other things, a whole range of words expressing the notion of 'time' in different guises (*tempo, momento, sosta, vacanza, durata*, etc.). The fact that *breve* is not predicated of *settimana*, and that *corto* is rather used in that context (cf. *settimana corta*) indicates: (1) that this is a typical case of 'frozen expression' or collocational use of *corto*, with a fairly idiosyncratic meaning, (2) that there is a potential link between the two areas of *corto* and *breve*, edging the semantic dimension of time, which failed to be represented in our diagram, presumably since it does not appear to be reinforced by other similar links in the context of 'smallness'.

Finally, Fig. 6 shows the bottom corner to the left of

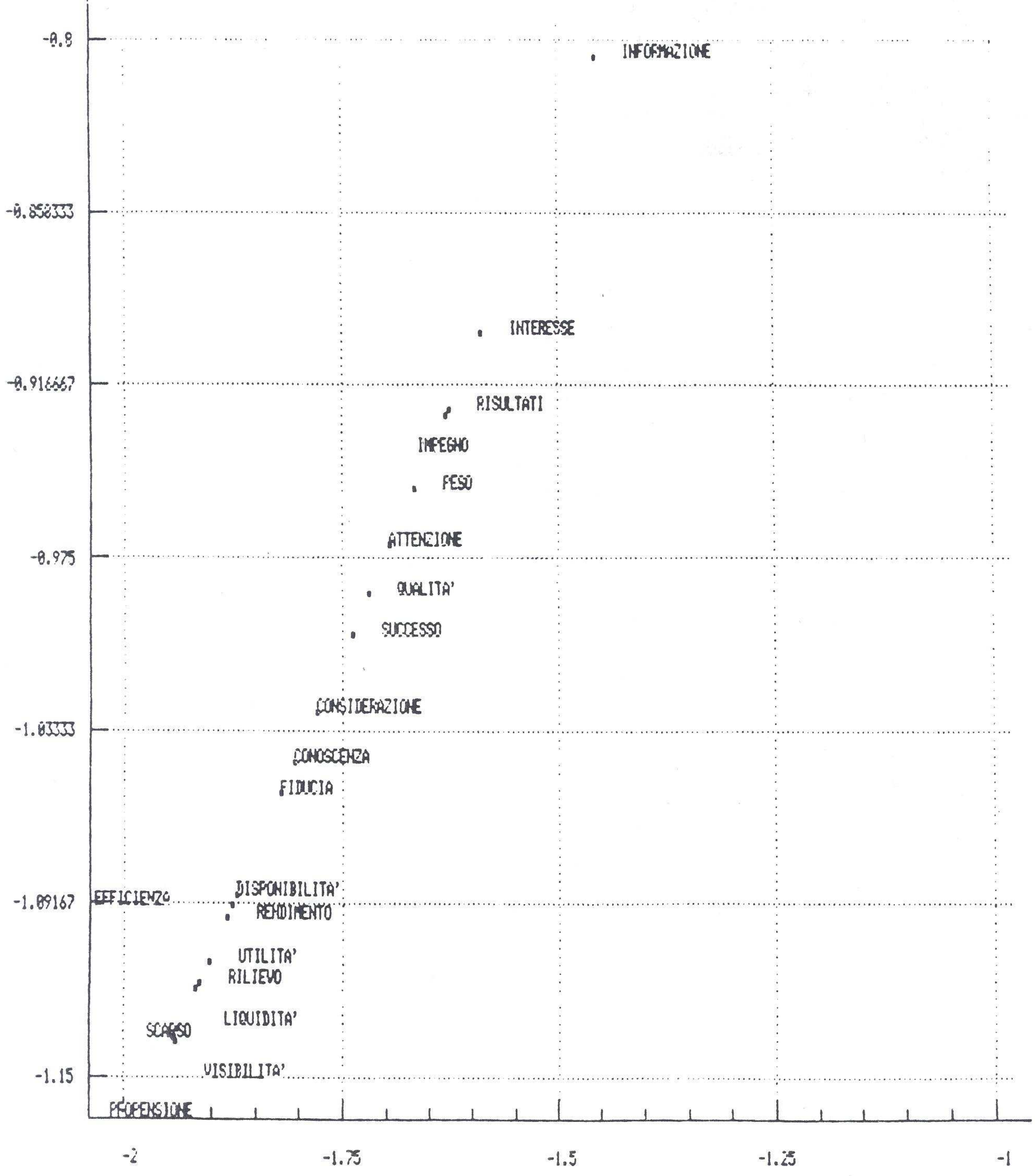

Fig. 6

the overall configuration. Here, *scarso* is accompanied by abstract nouns only, mostly derivatives. *Scarso* turns on the evaluation of quantity/size in given circumstances, by expressing that something is smaller than it should be. Our configuration suggests that, typically, the judgement expressed by *scarso* is predicated of nouns meaning attitudes (*propensione, attenzione, impegno, interesse, fiducia*, etc.), qualities (*utilita', disponibilita', visibilita', qualita'*, etc.), generic words referring to things from the point of view of their function or role, rather than by means of describing their own inherent properties (*liquidita', conoscenza, informazione*, etc.).

Note that, in the light of what observed so far, verticality in our diagram seems to carry with itself the semantic opposition±abstract: concrete nouns mostly concentrate in the top central area of the diagram; abstract nouns tend to be spread around at the bottom. Moreover, the latter are not 'sprayed' randomly. The notion of time, and what we might call 'length in relation to time' (cf. *distanza, racconto, descrizione*, as opposed to 'sheer' length as incorporated in *calzoncini, capelli, gonna*, etc.) is concentrated in the right corner; on the other hand, 'abstractness' as 'qualities', 'attitudes' and generic 'nouns' (cf. examples above) is represented to the left. The overall configuration is remarkably consistent and easily interpretable. Various levels of information concerning selectional restrictions, lexical collocations, and semi-idiomatic expressions are naturally represented along a continuum.

5.3 *Conclusions*

It is noteworthy that our lexical multidimensional charts are not neurobiological models of word interconnectivity in the mental lexicon, neither pictures of word learning. However dynamic our strategy is, we are more interested in the final stage of the process, than in the process itself. In the perspective of integrating textual information into a suitable computational lexicon, it makes no sense to stop halfway through in the process of extracting lexical knowledge from large text corpora: rather, the bigger the corpus being processed is, the better.

Moreover, as to the central problem of identification of word senses, the task of singling out different meaning shades is taken to be contingent on their pertinence: more figuratively, different slices of the multidimensional pie (the semantic hyper-space) carry with themselves a different bunch of word senses for the same word entry.

Polysemy and generality of meaning are no longer conceived of as alternative concepts shaping up our mental representation of word senses in fierce competition. Rather they are two complementary ways of accounting for lexical data and their structure. Discriminant analysis points to a clever idea: polysemy fades into generality when a semantic opposition (e.g. ±abstract) which used to be felt crucial for a certain range of purposes, has lost its relevance.

This leads immediately to our final observation: detailed differences of meaning between word senses are more likely to be dynamically generated according to varying conceptual standpoints, as they emerge from the embedding syntagmatic situation of their use, than to be, once for all, fixed within our mental dictionary, to be selected in the process of interpreting texts. Casting this type of 'fluid' lexical knowledge into the grid of an effective lexical database is still an open challenge for some time to come.

References

Atkins, B. T. and Levin (1988). Admitting impediments. *Information in Text: Proceedings of the 4th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, pp. 17–36.

Bindi, R., Monachini, M. and Orsolini, P. (1989). Italian Reference Corpus. ILC-CNR-3, Pisa.

Boguraev, B., Briscoe, E. J., Calzolari, N., Cater, A., Meijs, W. and Zampolli, A. (1988). Acquisition of Lexical Knowledge for Natural Language Processing Systems, (ACQUILEX), Technical Annex, ESPRIT Basic Research Action No. 3030, Cambridge.

Bortolini, U., Tagliavini, C. and Zampolli, A. (1971). Lessico di frequenza della lingua italiana contemporanea, Garzanti, Milano.

Brunet, E. (1983). Le style de Proust dans la Recherche du temps perdu. Etude quantitative. *Linguistica Computationale*, Vol. 3 suppl., Pisa.

Calzolari, N. and Bindi, R. (1990). Acquisition of lexical information from a large textual italian corpus. In *COLING '90 Proceedings*, edited by H. Karlgren, Helsinki, Vol. 3, pp. 54–9.

Calzolari, N., Cotoneschi, P. and Monachini, M. (1990). Translation Equivalences: English to Italian, In J. Sinclair (coord.), *The Prospect for a Multilingual Database*, Report to the Council of Europe.

Church, K. and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceeding of the 27th Annual Meeting of the Association of Computational Linguistics*, Vancouver, British Columbia, pp. 76–83.

Church, K., Gale, W., Hanks, P. and Hindle, D. (1990). Using Statistics in Lexical Analysis. AT&T.

Churchland, P. (1986). Some reductive strategies in cognitive Neurobiology. *Mind* XCV, n. 379, 279–309.

Cotoneschi, P. and Monachini, M. (1991). An empirical experience in the utilization of the Italian Reference Corpus in meaning analysis. In *Proceedings of ACH–ALLC '91 Conference*, edited by D. Ross and D. Brink, Tempe (Arizona) and forthcoming in *Research in Humanities and Computing*, Oxford University Press, Oxford.

Kelly, M., Bock, K. and Keil, F. (1986). Prototypicality in a Linguistic Context: Effects on Sentence Structure. *Journal of Memory and Language*, 25: 59–74.

Kelly, Stone (1975). *Computer recognition of English Word Senses*.

Monachini, M. and Calzolari, N. (1991). Translation Equivalences: English to Italian. 'Hear' and 'Listen', 'Calendar' and 'Diary', in J. Sinclair (coord), *Multilingual Lexicography Project*, Report to the Council of Europe (for the period up to 31 Maggio 1991).

Rieger, B. B. (1991). *On Distributed Representation in Word Semantics*. ICSI, Berkeley.

Rosch, E. (1975). Cognitive Reference Points. *Cognitive Psychology*, 7: 532–47.

Sinclair *et al.* (1990). The prospect for multilingual databases. A report to the Council of Europe. Birmingham.

Zampolli, A. (1990). *Project definition for the constitution of a network of European textual reference corpus.* Pisa.