Acquisition of Lexical Information

from a Large Textual Italian Corpus

Nicoletta CALZOLARI - Remo BINDI

Istituto di Linguistica Computazionale del CNR, Pisa
Dipartimento di Linguistica, Universita' di Pisa

Via della Faggiola 32
56100 PISA - ITALY
e-mail: GLOTTOLO @ ICNUCEVM.BITNET

1. Introduction: LDB requirements

The creation and development of a large Lexical Database (LDB) which, until now, mainly reuses the data found in standard Machine Readable Dictionaries, has been going on in Pisa for a number of years (see Calzolari 1984, 1988, Calzolari, Picchi 1988). We are well aware that, in order to build a more powerful LDB (or even a Lexical Knowledge Base) to be used in different Computational Linguistics (CL) applications, types of information other than those usually found in machine readable dictionaries are urgently needed. Different sources of information must therefore be exploited if we want to overcome the "lexical bottleneck" of Natural Language Processing (NLP).

In a trend which is becoming increasingly relevant both in CI proper and in Literary and Linguistic Computing, we feel that very interesting data for our LDBs can be found by processing large textual corpora, where the actual usage of the language can be truly investigated. Many research projects are nowadays collecting large amounts of textual data, thus providing more and more material to be analyzed for descriptions based on measurable evidence of how language is actually used. We ultimately aim at integrating lexical data extracted from the analysis of large textual corpora into the LDB we are implementing. These data refer, typically, to:

i) complementation relations introduced by prepositions (e.g. *dividere* <divide> subcategorizes for a PP headed by the preposition *in* <in> in one sense, and by the preposition *fra* <among> in another sense);

ii) lexically conditioned modification relations (*una macchina potente* <powerful car>, *un farmaco potente* <potent medicine> and not *forte* , while *un caffe' forte* <strong coffee>, *una moneta forte* <strong currency> and not *potente* <powerful>);

iii) lexically significant collocations (*prendere una decisione* <to take a decision> and not *fare una decisione* <to make>, *prestare attenzione* <to pay attention> and not *dare* <to give>);

iv) fixed phrases and idioms[1] (*donna in carriera, dottorato di ricerca, a proposito di*);

v) compounds (*tavola calda, nave scuola*).

All these types of data are a major issue of practical relevance, and particularly problematic, in many NLP applications in different areas. They should therefore be given very large coverage in any useful LDB, and, moreover, should also be annotated, in a computerized lexicon, for the pertinent frequency information obtained from the processed corpus, and obviously updated from time to time. As a matter of fact, dictionaries now tend to encode all the theoretical possibilities on a same level, but "if every possibility in the dictionary must be given equal weight, parsing is very difficult" (Church 1988, p.3): they should provide information on what is more likely to occur, e.g. relative likelihood of alternate parts of speech for a word or of alternate word-senses, both out of context and if possible taking into account contextual factors.

Statistical analyses of linguistic data were very popular in the '50s and '60s, mainly, even though not only, for literary types of analyses and for studies on the lexicon (Guiraud 1959, Muller 1964, Moskovich 1977). Stochastic approaches to linguistic analyses have been strongly reevaluated in the past few years, either for syntactic analysis (Garside et al. 1987, Church 1988), or for NLP applications (Brown et al. 1988), or for semantic analysis (Zernik 1989, Smadja 1989). Quantitative (not statistical) evidence on e.g. word-sense occurrences in a large corpus have been taken into account for lexicographic descriptions (Cobuild 1987).

---

[1] Here and in the following we have not translated idiomatic phrases and compounds, because there is no point in giving the literal translation of the single words.

The claim of this paper is that the above types of linguistic information (i-v), to be made available for our LDB, can be partially extracted by processing and analyzing very large text corpora, with quantitative/statistic methods.

2. The Italian Reference Corpus

The corpus (see Zampolli 1988) on which we are now conducting our analysis is being produced by the ILC and an Italian publishing house (see Bindi et al. 1989). The project was begun in 1988. The corpus now contains about 12 million words, and the first goal is to reach 20 million words by the end of '90. When completed, the corpus will be balanced among journals, novels, manuals, scientific texts, 'grey' literature, etc. The corpus is presently unproportioned, because we first processed and inserted up to about 8 million words from journals, newspapers, magazines, etc., while we are now inserting data from novels and from the scientific and technical literature.

The present study is conducted on the first section of the corpus, but we obviously intend to extend the analysis to the other sections as soon as they become available.

We describe two types of quantitative analyses whose aim is to extract information on:
a) the *strength of association* between two words;
b) fixed phrases or idioms.

3. The *strength of association* within word-pairs

As regards the first point we have used the method of measuring the *association ratio* between two words as described by Church and Hanks (1989). The value of the association ratio reflects the strength of the bond between the two words taken into account. The method is very simple. The association ratio between any two words x and y appearing together in a window of five words in the corpus is based on the concept of "mutual information" defined as:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)\,P(y)}$$

where P is the probability. We refer to Church and Hanks (1989, pp. 77-78) for a detailed explanation of the formula and of how the association ratio slightly differs from it, given that we are more interested here in the linguistic and lexicographic evaluation of the numerical results deriving from its application.

In addition to this we have introduced the measurement of the so-called *dispersion*, in order to obtain - linked to the association ratio - quantitative information on the distribution of the second word of the word-pair in the selected window. We wanted in fact to complete the simple frequency notion for a word-pair with that of frequency stability or dispersion, i.e. to add to the frequency a measure of how it is distributed over the different positions of the window. In this way we evaluate the uniformity of repartition of frequency of the second word over the considered span. We have used the formula described in detail in Bortolini et al. (1971, pp. 23-31), even though used here for different purposes.

We give some examples in Table 1, where f(x,y) is the frequency of occurrence of words x and y together and in this order in a window of 5 words, gap is equal to the number of words between x and y (if gap = 0 then x and y are immediately adjacent), f(x) and f(y) are the frequencies of occurrence of x and y independently in the corpus, ass.ratio is the result of application of the formula to x and y, dispersion calculates how the second word is distributed within the considered window.

This last information is very useful not only to evidence words belonging to fixed phrases, but especially while trying to evidence syntactic relationships. If the dispersion is 0 or near to 0, all or most of the occurrences of the second word are concentrated in the same position. This means that the position and distance of the two words is always the same, and it is therefore a strong measure for evidencing "fixed phrases" or "compounds" with no variation inside. When viceversa its value approaches 1, y is almost equally distributed in the four positions of the considered span. Thus, the combination of a not very high (but above a certain level) ass.ratio with dispersion values near to 1 is more typical of syntactic types of collocations, giving e.g. information on prepositional government.

We wish to highlight here some of the results achieved by the application of these statistical measures to the Italian corpus, and mainly to evaluate their linguistic relevance.

Table 1.

		f(x,y)	gap = 0	gap = 1	gap = 2	gap = 3	f(x)	f(y)	ass.ratio	dispersion
Stati	*Uniti*	2047	2042	0	0	5	5850	2159	10.34	0.003
punto	*vista*	832	0	831	0	1	4396	1974	9.58	0.002
opinione	*pubblica*	272	272	0	0	0	657	1315	11.30	0.000
presta	*a*	33	14	9	6	4	85	120969	4.68	0.736
spendere	*per*	36	8	8	9	11	183	101862	3.95	0.921
e'	*ambizioso*	20	5	5	5	5	115476	123	3.49	1.000

From the present corpus of 8,032,667 occurrences (tokens) and 178,811 different word-forms (types), we obtained 26,473,263 word-pairs (tokens) in a window of 5 words (and not 32,000,000, as the window is not extended beyond any strong punctuation mark) and 8,716,446 different word-pairs (types). After discarding all the pairs with f(x,y) < 4, because they were too rare and of no linguistic relevance, 787,878 word-pairs were obtained, which were eventually reduced to 322,718 after eliminating those with association ratio < 3 (the pairs seem to be linguistically irrelevant below this level).

We must also recall that the data to which we have applied our measures are articles from many different types of newspapers, journals, etc. - i.e. many short texts - , so that there is no bias towards clustering tendencies of words such as could appear in longer texts, like entire novels.

If we order the word-pairs by decreasing value of the *association ratio*, and examine the types of word combinations appearing in the different positions of the file, we observe a different typology of word combinations according to the different levels:

i) at the top;
ii) in the center;
iii) towards the lower interesting values, which for Italian seems to be a little higher than for English, i.e. around 3.5;
iv) below this significant value, until reaching the few negative values.

For example at the top, i.e. with very high values (ranging from 22.93 to about 15), we find the following categories of word-pairs:

- proper nouns, titles, etc. (e.g. *Oci Ciornie 20.6, Cyrano Bergerac 20.1, Montgomery Clift 20.1, Ursula Andress 19.9*);
- foreign (usually English) compound words or fixed phrases (*value added 19.8, pax Christi 17.7, teen ager 17.7, drug administration 17.3*);
- Italian compounds of words belonging to specialized languages, which almost never occur in everyday language (*bismuto colloidale 20.1, tomografia assiale 19.8, marmitte catalitiche 19.6, nitrato ammonio 19.5, accoppiatore acustico 17.5*);
- co-occurring technical words, which again appear very rarely (*laringiti tracheiti 20.3, idrologia climatologia 20.2, capperi cetriolini 19.6, prefetti questori 18.5, antisettiche antispasmodiche 17.8*);
- fixed phrases or idioms whose component words are not frequent in ordinary language (*volente nolente 20.6, specchietto allodole 18.8, bla bla 18.00, batter ciglio 17.2, cartoni animati 16.5, spron battuto 15.5*);
- modification relations between low frequency Adjectives and Nouns (*sostantivi plurali 19.9, forbicine affilate 18.4, gradazione alcolica 18.1, giubbotti antiproiettile 17.4, salmone affumicato 17.1*);
- modification relations between Noun and Noun of a PP, both of low frequency (*cartina tornasole 18.3, filetti alici 17.7, siepi bosso 15.9, spicchio aglio 15.5*).

These word-pairs share the following properties: both the words are of very low frequency, and almost always appear only together in the same context.

The characteristics of the different types of combinations appearing within the other ranges of the association ratio value, i.e. from ii) to iv) above (for example, at the value levels when more specific grammatical/syntactic information appears), are very different and present quite interesting properties.

Thus, we have observed how the measure of the association ratio gives quantitative/statistical evidence to a number of lexical, syntactic and semantic relationships between word-pairs. These relationships are essential for codification in an LDB, and cannot be achieved with the same "objectiveness", and certainly not to the same extent, by other means such as e.g. lexicographers' intuition.

Among the syntactic relationships, particularly relevant is the data which regards the prepositions marking the different arguments of verbs, adjectives and nouns, together with their relative frequency. This is very important information to be inserted in the LDB (especially of Italian), provided we have no dictionary source for this type of complementation as for example the Longman dictionary for English. Other syntactic data concern the type of sentential complementation, mainly for the verbs.

We notice, for example, that in all their inflections the verb *rischiare* <to risk> and the noun *rischio* only subcategorize for the preposition *di* <of>; the same holds for the adjective *capace* <able>. This information is simply a confirmation of their only possible prepositional complementizer. The verb *pensare* <to think> is found with *a, che, come, di* <to, that, how, of>, i.e. with all its theoretical possibilities of prepositional and sentential government, while *parlare* <to speak> is more frequently associated only with *con, di* <with, about>, and not with *a* <to>, which should be found in principle. *Dividere* is mostly associated with *con, da, in* <with, from, in>, and not with *tra* <among>. These quantitative data can be associated to the different subcategorization frames and can be helpful for complementation rules, to decide on ambiguous attachments of PPs.

As a next step, we are trying to correlate the different complementation patterns evidenced by some word-pairs with other lexical information (found in the environment of these first word-pairs) which can be used as a clue for semantic disambiguation. For example, if we take the word-pairs *dividere con, dividere da, dividere in*, we must look at the surrounding context and see which generalizations can be done at the semantic level for

the three types of subcategorization. These may in fact correspond to different word-senses.

Very useful data of both syntactic and lexical/semantic relevance concern the so-called *support verbs* (see Gross 1982) for Nouns (usually deverbal or Action nouns) or for Adjectives. We observe for example:

compiere accertamenti	*10.8*
fare affidamento	*8.1*
avere accesso	*5.3*
condurre/effettuare analisi	*8.3/7.3*
avuto accoglienza	*8.0*
prendere decisione	*9.7*
rendere accettabile	*8.9*
rendere accessibile	*9.4*

This sort of information on support verbs is of essential importance for language generation (see Mel'cuk, Polguere 1988), and cannot be predicted in any other way, but can only be given either by observation or by introspection. The automatic collection of these data is thus an important shortcut towards their extensive coverage in a LDB. Their semantics can be rather easily inferred by the type of support verb (there is a finite list of them) and given by rule.

Purely semantic data mainly regard typical collocations, e.g. between Adjective and Noun (see below), or between Verb and Adverb, or between Verb and typical Subjects and/or Objects (*fondare colonia 11.4, abbassare colesterolo 11.3, distogliere attenzione 10.9, attirare attenzione 10.7, prestare attenzione 10.5, sparo' colpo 10.6*).

Interesting data are also found concerning the semantic field of certain words, and obviously words belonging to a fixed phrase. For co-occurrences of Nouns belonging to the same semantic field an example is:

abbigliamento accessori	*9.6*
abiti accessori	*9.4*
borse accessori	*9.3*
scarpe accessori	*9.0*

Examples of fixed and/or idiomatic phrases are:

battuta arresto	*11.7*	*(battuta d'arresto)*
polmone acciaio	*11.6*	*(polmone d'acciaio)*
primo acchito	*10.1*	*(di primo acchito)*

As this method is only used to work on couples of words, it is clear that we do not generally obtain the whole phrase. It is for this reason that we have developed, especially for this type of data, other quantitative tools which are described in section 4, whose results will supplement those provided by this method.

A number of different observations can be made for the word-pairs, according to whether they are sorted on the right or the left word. If we examine the left contexts (i.e. if words are ordered on the right), we are more likely to gather information on e.g. the Nouns which are typically modified by a given following Adjective (*sorriso accattivante 11.3; luce accecante 10.8; luce accesa 8.7, radio accesa 9.7, colori accesi 10.0, toni accesi 11.2, forno acceso 10.7, fuoco acceso 8.5*). If vice-versa we examine the right contexts, it is easier to collect data on the Nouns which are typically modified by a given preceding Adjective (*costante aumento 7.6, costante contatto 6.4, costante miglioramento 7.9, costante riferimento 7.4, costante temperatura 8.1*).

In the left contexts again we find together data which regard which Adjectives are typical pre-modifiers of a given Noun (*forte accento 8.6, inconfondibile accento 12.0; difficile accesso 5.3, facile accesso 5.7, libero accesso 7.5; buona accoglienza 8.7; antico amore 4.8, buon amore 3.4, eterno amore 7.0, grande amore 5.2, improvviso amore 5.5, ultimo amore 4.4, vecchio amore 3.7, vero amore 3.7*), or which types of Nouns are the governors of PPs with a given Noun as head (*controllo armamenti 8.9, limitazione armamenti 11.9, riduzione armamenti 9.1, settore armamenti 6.9*).

When analyzing the left contexts, we also find high association ratios for certain types of grammatical structures such as: compound verbs (with *essere* <to be> or *avere* <to have> as left word), reflexive or intransitive pronominal verbs (with the particle *si* on the left), reciprocal verbs (with the particles *ci, vi*), etc. All these types of data are obviously important for the creation of an exhaustive LDB.

As a final remark we can add that it would certainly be useful to make the same calculations on a tagged (for POS) corpus, in order to obtain relevant information for the lemmas; however, we must observe that different word-forms of the same lemma often present very different combinatorial properties, both at the grammatical/syntactic level and at the lexical/semantic level. When compacting information for a single lemma we must therefore be careful not to lose data which are relevant to particular inflected forms. This kind of information is again particularly important for practical NLP applications.

4. Fixed phrases and idioms

Mainly for the detection of "stereotypes" in texts we have implemented and are now refining other quantitative/statistical tools not limited to couples of words.

In order to collect data concerning specifically fixed phrases or multi-word units, we first calculated the frequency of occurrence in the corpus of all identical couples, triples, and so on, up to seven-word syntagms.

Also for this data we calculated the *dispersion*, and we also calculated the so-called *usage*. Also usage is defined according to Bortolini et al. (1971) as: U = FD, i.e. Usage equal to Frequency by

Table 2.

	'85	'86	'87	'88	Total	Disp.	Usage	(Novels)
per la prima volta	136	119	111	123	489	0.96	468.13	102
dal punto di vista	64	76	93	77	310	0.92	286.20	21
in tutto il mondo	73	78	60	66	277	0.94	261.22	2
un vero e proprio	43	23	21	25	112	0.82	91.74	29

Dispersion. It is therefore equal to Frequency when the word is uniformly distributed in the different years (and genres), and is equal to 0 when Dispersion is 0, i.e. if all occurrences were concentrated in a single year (or genre). Usage is as nearer to Frequency as much the distribution is uniform, and decreases proportionally while Dispersion is decreasing.

In this case dispersion and usage were first calculated on the sections of the corpus which refer to the 4 years of publication of the journals (from 1985 to 1988), in order to point out, among others, the appearance (or disappearance) of phrases, compounds, and stereotypes in general. We then compared a subset of all the press data with a subset of novels of analogous size, and again calculated dispersion and usage in order to evidence eventual difference of distribution of these fixed phrases between press and novels.

The data (of the two types) were then sorted in different ways: by alphabetical order of the n-tuples, by frequency of occurrence of the n-tuples, by dispersion, by usage. From each ordering we gather data which can be used in a variety of ways or can evidence different types of phenomena. An example at the beginning of the file of the quadruples ordered by usage (in decreasing order) is found in Table 2 (with figures for dispersion and usage only concerning press data, i.e. the first four columns; the column for Novels, of the same size as each year column, has been inserted in the table from the second comparison just for curiosity).

The data, i.e. all the n-tuples of different lengths, were also merged in a single file, to evidence the precise length of each given phrase. For example, *vero e proprio* is in a very high position for its frequency in the set of triples, but the fact that *un vero e proprio* is also in a very high position in the set of quadruples means that this is the size of the 'true fixed phrase'. Other observations on the linguistic results evidenced by this method will be made in the presentation.

5. Final remarks

In the next months we intend to experiment with other statistical formulas (e.g. those used by Smadja and Choueka) on the corpus (which will also contain the novels and other types of texts).

The first stage of the research consists in a careful linguistic analysis of the results obtained by the different statistical tools we are now implementing and applying. By this analysis - performed according to different parameters, both from the statistical and linguistic/lexicographic viewpoints - we aim at achieving a twofold objective. On the one side we aim at setting up the beginning of a sound methodology to semi-automatize the extraction of at least part of the relevant syntactic/semantic relationships from the corpus; on the other side we hope we shall be able to build a model of the "actual" modification and complementation relations (out of the theoretical a-priori possibilities), of the "actual" lexical collocations, of the "actual" stereotypes in the Italian language.

One of the claims of this project is that the linguistic information embodied in all these quite different types of lexical collocations - once they have been supplied in a systematic way by a computational lexicon which is also annotated for frequency - can be helpful for lexical disambiguation in analysis and crucial for lexical selection in generation. Our method should be seen as a strategy to obtain in a semi-automatic way, and for a large portion of the lexicon, a formalization of many of the types of lexical relations coded, for example, in the Mel'cuk lexicon. This should be an enhancement both for a more concrete and objective lexicography (the results will be in fact evaluated in the next months in a true lexicographic environment), and for a more comprehensive and "data-based" linguistics.

Acknowledgment

We wish to thank our Referees for useful comments and suggestions, and A. Zampolli for helpful discussions.

References

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

Bortolini, U., C. Tagliavini, A. Zampolli, (1971), *Lessico di Frequenza della Lingua Italiana Contemporanea*, Garzanti, Milano.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P., (1988), "A statistical approach to language translation", *Proceedings of the 27th International Conference on Computational Linguistics*, Budapest.

Calzolari, N., (1984), "Detecting patterns in a Lexical Database", *Proceedings of the 10th*

International Conference on Computational Linguistics, Stanford (CA), 170-173.

Calzolari, N., (1988), "The dictionary and the thesaurus can be combined", in M. Evens (ed.), *Relational Models of the Lexicon*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, 75-96.

Calzolari, N., (1989), "Lexical Databases and Textual Corpora: perspectives of integration for a Lexical Knowledge Base", *Proceedings of the 1st International Lexical Acquisition Workshop*, Detroit, Michigan.

Calzolari, N., E. Picchi, (1988), "Acquisition of semantic information from an on-line dictionary", *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary, 87-92.

Choueka, Y., (1988), "Looking for needles in a haystack", *Proceedings of the RIAO*, 609-623.

Church, K., (1988), "A stochastic parts program and noun phrase parser for unrestricted text", *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing*.

Church K., P. Hanks, (1989), "Word association norms, mutual information and lexicography", *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 76-83.

Cobuild, (1987), *Collins Cobuild English Language Dictionary*, Collins, Glasgow.

Garside, R., Leech, G., Sampson, G., (1987), *The Computational Analysis of English - a corpus based approach*, Longman, London.

Gross, M., (1982), "On the notion of support verb", seminar at the Simon Fraser University, B.C. Canada.

Guiraud, P., (1959), *Problemes et methodes de la statistique linguistique*, D.Reidel, Dordrecht.

Hindle, D., (1989), "Acquiring disambiguation rules from text", *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 118-125.

Mel'cuk, I., Polguere, A., (1988), "A formal lexicon in Meaning-Text theory", *Computational Linguistics*, 13(3-4).

Muller, Ch., (1964), *Essai de Statistique Lexicale*, Klincksiek, Paris.

Muller, Ch., (1965), "Frequence, dispersion et usage: a propos des dictionnaires de frequence", *Cahiers de Lexicologie*, II, 32-42.

Moskovitch, W.A., (1977), "Polysemy in natural and artificial (planned) languages", *SMIL, Journal of Linguistic Calculus*, Skriptor, 1, 5-28.

Smadja, J., (1989), "Macrocoding the lexicon with co-occurrence knowledge", *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, Michigan.

Webster, M., M. Marcus, (1989), "Automatic acquisition of the lexical semantics of verbs from sentence frames", *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, 177-184.

Zampolli, A., (1988), "Progetto Strategico *Metodi e strumenti per l'Industria delle Lingue nella collaborazione internazionale*", ILC-CNR, Pisa.

Zernik, U., (1989), "Paradigms in lexical acquisition", *Proceedings of the First International Lexical Acquisition Workshop*, Detroit, Michigan.