OBSERVATION AND GENERALISATION: CORPUS-BASED LINGUISTIC ANALYSIS OF ITALIAN SPEECH ACT VERBS

Calzolari, Nicoletta

Istituto di Linguistica Computazionale del CNR, Via della Faggiola 32, 56100 Pisa, Italia, Tlfno: 07-39-50560481, e-mail: glottolo@ilc.pi.cnr.it

(Recibido Febrero 1998; aceptado Abril 1998)

BIBLID [1133-682X (1997-1998) 5-6; 59-73.]

Resumen

Presentamos algunas observaciones sobre la lengua italiana dentro del proyecto LRE.DELIS en el marco de la UE. Se aplican dos hipótesis complementarias como medio para descubrir la correlación entre aspectos sintácticos y semánticos - una aproximación lexicográfica basada en corpus y otra de análisis basado en los marcos semánticos.

Palabras clave: basado en corpus, lexicografía, basado en marcos, semántica, lexicón, verbos.

Abstract

I present some observations on the Italian language within the framework of the EC funded LRE project DELIS. Two complementary hypotheses - a corpus-based lexicographical approach and a frame semantic analysis - are used as a means to discover the correlation between syntactic and semantic aspects.

Key words: corpus-based, lexicography, frame-based, semantics, lexicon, verbs.

Résumé

Nous présentons quelques remarques sur la langue italienne dans le projet LRE DELIS dans le cadre de la UE. Nous appliquens deux hypothesès complémentaires comme myens pour découvrir la corrélation entre les aspects syntaxiques et sémantiques: une approche léxicographique fondée dans des corpus et une analyse fondée sur des structures sémantiques.

DOI: http://dx.doi.org/10.25267/Pragmalinguistica.1997.i5.03

Mots-clés: fondé sur des corpus, léxicographique, fondé sur des structures, sémantique, léxique, verbes.

Sumario

1. Theoretical framework: corpus based lexıcography and frame semantics. 2. Corpus analysis of semantic classes of verbs. 3 A few observatıons on speech act verbs. 4. Correlation between meaning and morphology 5 Syntactic properties and correlation between semantics and syntax. 6. Corpus evidence vs. introspection as a basis for lexıcographıc description: actual vs. potential. 7. Conclusions.

1. Theoretical Framework: Corpus based Lexicography and Frame Semantics

I present here some observations stemming from work done on the Italian language, in the framework of the EC funded LRE project DELIS[1], on "methods and tools for corpus-based lexicography".

Two hypotheses - a methodological and a theoretical one - form the background for the present observations, and for DELIS work in general: i) a corpus based lexicographical approach, and ii) a frame based semantic analysis.

i. We assume that only a careful and detailed analysis of corpus data can provide a sound basis for a realistic approach to lexicon building, be it a human oriented dictionary or a computationally oriented lexicon for language engineering applications.

ii. Obviously corpus data cannot be used in a simplistic way. In order to become usable, the data must be analysed according to some theoretical hypothesis, on the basis of which to model and structure what would otherwise be an unstructured set of data. The best combination of both the empirical and the theoretical approach is when the theoretical hypothesis itself emerges from and is guided by successive analyses of the data, and the encoding of lexical data is cyclically refined and adjusted to textual evidence.

[1] LRE DELIS- funded by DGXIII E of CEC- was coordinated by University of Stuttgart, and partners were Universities of Amsterdam (VUA), Clermont-Ferrand, Copenhagen, Istituto di Linguistica Computazionale of CNR (Pisa), Linguacubum Ltd (London), and publishing houses Den Danske Ordborg (Copenhagen), Oxford Universıty Press and Van Dale Lexicographie (Utrecht).

Within the DELIS project we have used, as a theoretical modelling hypothesis, the frame semantics approach defined by Fillmore (see e.g. Fillmore and Atkins 1992) as a guide in the analysis of textual data, and in the subsequent design of the lexical entry. An essential descriptive strategy of this framework links together the semantic and syntactic descriptive levels. An important characteristic of this approach to corpus analysis and to lexicon building is that it aims at the reusability of its results in the context of NLP applications. Moreover, when deciding on the methodology for the analysis of the corpus, this approach allows the lexicographer/linguist to either i) start with an intuitive semantic characterisation of the so-called semantic "frame elements" and only afterwards look for their syntactic realisation or, vice versa ii) start with a description of the main syntactic constructions and only afterwards give a semantic interpretation in terms of frame elements.

The two methodological approaches are complementary and can be chosen according to particular characteristics of the semantic class of words being analysed by the lexicographer (see Alonge et al. 1993).

In this paper we will not dwell upon the relevance of different methodologies of corpus analysis according to different types of semantic classes of words, but will provide a few examples of the linguistic results of a corpus based lexical analysis. In doing so particular attention will be paid to the correlations between different levels of linguistic description. The DELIS project focused on the correlation between syntactic and semantic aspects, but it is evident that other linguistic aspects - such as morphology, morphosyntax, lexical co-occurrence, collocational data, etc. - are closely interrelated, and these relations have to be captured when designing a lexical entry, in particular when accounting for the phenomenon of meaning discrimination. It is precisely the complexity of the interrelationships between all these aspects which makes semantic disambiguation such a difficult task in NLP. One of our aims is to use textual corpora as a means to discover and reveal the intricacy of these relationships, and frame semantics as a means to unravel and disentangle this complex situation into elementary and computationally manageable pieces.

2. Corpus analysis of semantic classes of verbs

The first step taken within the DELIS project was to collect, in a systematic way, from actual language use as can be found in text corpora[2], evidence of regularities (or

[2] For analysis of Italian we made use of the Italian Reference Corpus,built at ILC in cooperation with Elemond (see Bindi, Monachini, Orsolini, 1991), at the time of analysis composed by about 15 Millions

idiosyncrasies) in the behaviour of a few verb classes at different linguistic levels (morphosyntactic, syntactic, semantic, and the interrelationships between these levels).

The results of this systematic analysis, and the generalisations allowed by the data within the theoretical framework adopted as a working hypothesis, served as a starting point for the next phase of design and population of the corresponding lexical entries (see Monachini et al. 1994) in the computational lexicon.

The following procedure - described here in a very schematic way - was used.

A so-called "DELIS Encoding Scheme" (see Krueger and Heid 1993) was devised, as a very detailed guideline for corpus annotation (with respect to morphosyntactic, syntactic, semantic, collocational, etc. information), with examples from each language of the project (Dutch, English, French, German, Italian), and with descriptions of their peculiarities vis-à-vis the general guidelines for each language, in order to take into account particular requirements of each specific language, while maintaining a common encoding scheme for the project.

The corpus was automatically and/or manually annotated using the Delis Encoding Scheme. The results of the linguistic analysis of the annotated corpus have highlighted, in particular, the importance, for the lexicographical description of a lexical entry, of many types of correlations between different levels of linguistic analysis (morphology, syntax, semantics), and have provided the basis for the subsequent generalisation phase. Problems in encoding have also been evidenced.

Different types of basic regularities in the collected data were systematically recorded, to be considered further when modelling the relevant lexical entries (see Monachini et al. 1994).

A number of Perception and Speech Act verbs were analysed. The more general verbs in the field were selected, aiming at a representation of the field itself. Other more specific verbs were chosen because they presented different characteristics with respect to the general ones. The same verbs were selected within each language, thereby allowing for a cross-linguistic comparison.

Lists of the inflected word-forms occurring in the Corpus (for the Italian Corpus see Bindi et al. 1991) for the chosen verbs were produced, together with their frequency of occurrence (this refers to the character string, therefore for homograph word-forms this number may refer to word-forms belonging to different parts-of-speech, e.g. '*promessa*' can

occurrences from a variety of text types.

be either the Past Participle Feminine Singular of the Verb or the Noun Feminine Singular). Homograph forms were marked and disambiguated.

In the analysis, only verb occurrences were considered, and - as a general DELIS decision - those verb occurrences which were a part of a Multiword Expression, better considered in the lexicon as an independent Lexical Unit, were put aside. The Multiwords encountered in this analysis were simply listed, without attempting any further distinction among different types of Multiwords.

The first, obvious observation made was that the two selected semantic classes demonstrated very different semantic and syntactic behaviours, implying that a somewhat different approach to their lexicographical analysis and to corpus annotation may be more appropriate for the two considered fields.

We briefly outline below a few examples of the types of results of this analysis, focusing on one semantic class, i.e., Speech Act verbs.

3. A few observations on Speech Act verbs

From an analysis of the occurrences of all the inflected word-forms of speech-act verbs, we were able to capture:

1) Data which are sometimes - or often - neglected by traditional lexicography, e.g. i) meanings of the verb which are strictly linked to particular contexts or to particular morphological inflections, ii) many idiomatic expressions and collocations, iii) all the syntactic patterns allowed by a verb and their respective frequencies, iv) preferences given by a verb to particular syntactic constructions, lexical items, etc., v) evidence of syntactic, semantic and morphological clues which seem to be crucial in (semi-) automatic sense disambiguation and in characterising unambiguously a given use of the verb.

2) Data which could be taken into consideration in view of the future automation of corpus annotation and of knowledge extraction (e.g. syntactic patterns, verb arguments, etc.) from text corpora.

Indeed, the first phase of the detailed analysis of corpus occurrences was meant to highlight the possible syntactic configurations allowed by the verb and their correlation, if any, with the meanings of the verb and with the frame elements. Encoding and extraction of such data was done manually in DELIS through the use of a simple editing tool (see Federici 1993), but with the view of collecting evidence for its future automation.

Two facts immediately became apparent in the analysis of e.g. *dire* (to say) and *promettere* (to promise):

1) they were not overly complex from a semantic point of view, i.e. they did not present a very high degree of polisemy (not considering their appearance in Multiwords or in idioms to be headed as separate lexical entries);

2) they occurred, however, with a rather large variety of subcategorisation patterns, but only few of these patterns (very restricted in number) were univocal signs of a difference in meaning. Most of the patterns can be considered as different possible syntactic manifestations of the same basic meaning, maybe within a partially different template of surrounding Frame Elements.

The basic type of information that we can extract from the Corpus, for this class of verbs, is therefore of a syntactic nature, in particular phrase structure information plus grammatical relations and control information. This type of syntactic information was extracted and encoded by hand, but this type of annotation could be - at least partially - handled in an automatic way.

As already stated, the main type of variation encountered in the analysis of corpus occurrences of Speech Act verbs concerned their syntactic surrounding. It seemed, therefore, appropriate to annotate corpus occurrences according to the different types of syntactic environment. As a result of this syntactic annotation a list of syntactic phrasal patterns was produced, containing all the patterns actually occurring in the analysed Corpus subset. This list is not meant to be exhaustive, and will be updated when analysing other verb occurrences. It is being used as a basis for uniform, syntactic encoding and - most importantly - can be taken into account by a tool when searching, in a Corpus, for relevant surface syntactic patterns.

A field, containing a label for an element from this list, is therefore an important feature of the general DELIS Encoding Scheme. For example, to the word-form *disse* (told/said) in the context:

LC (Left Context) - *Il contadino mi*
(The farm-worker ... to me)

KW (KeyWord) - ***disse***
(told)

RC (Right Context) - *di aver fatto la guerra del 1716...*
(to have fought in the 1716 war)

we could associate a Syntactic Pattern code such as:

"IO+diInf(sc)" (ind.obj. + diInf. (subj.control)).

This value is selected by the annotator from a predefined list, or is assigned by a program after a phrasal analysis of the sentence.

4. Correlation between meaning and morphology

In a remarkable number of inflected word-forms, and sometimes in connection with particular syntactic patterns, the verb *dire* acquires a very specific meaning, not found with other inflections. In a multilingual context, these inflected word-forms require very specific translations which, moreover, do not necessarily always involve a speech-act verb in the target language.

This is the case with *diresti* (2nd Singular Conditional Present) (you would say), which is found in many occurrences in the following two constructions:

*Cosa\che ne **diresti*** + diInf. (subj.control) ?

*Cosa\che ne **diresti*** + PP(di) ?

with the meaning of *ti piacerebbe ?* (would you like ?).

Two examples are:

*Cosa ne **diresti** di andare a mangiare qualcosa ?*
(lit.: What would you say to go out and eat something ?)

*Che ne **diresti** di una fetta di focaccia.. ?*
(lit.: What would you say of a slice of plain pizza.. ?)

The same meaning should, in principle, be found also with the 3rd Singular Conditional Present *direbbe* (he would say) when it is used as the 3rd person polite form. We did not find this to be the case in our Corpus, but one should make use of a Speech Corpus to check whether this principle holds true.

The form *dici* (2nd Singular Indicative Present), when occurring in the same syntactic pattern, conveys a different meaning, i.e. "*cosa ne pensi?*" (what do you think about?):

*Che ne **dici** di questa storia?*
(lit.: What do you say of this fact?)

Another example of an idiosyncratic meaning, linked to a particular inflection, is provided by *direi* (1st Singular Conditional Present) (I would say), which in many occurrences is used in a parenthetical construction, usually, but not always, before an adjective. In these cases this word-form behaves, both at the semantic and syntactic levels, as an adverb, and acquires the meaning of the adverbial Multiword *per cosi' dire* (as it were).

Following are some examples:

*lo sguardo era sincero, **direi** amichevole*
(his glance was sincere, I would say friendly)

*in modo **direi** scorretto*
(in a way, I would say, incorrect)

*non lo fa trascinare, ma **direi** arrancare*
(it doesn't make him drag, but I would say struggle along)

*tre o quattro veicoli, **direi** camion o autobus*
(three or four vehicles, I would say lorries or buses)

This same construction is found in the 1st Plural Indicative Future *diremo* (we shall say), as in:

*l'iniziativa **diremo** reaganiana di Bush*
(the initiative, we shall say Reaganite, of Bush)

The fact - frequently encountered - that some syntactic and/or semantic behaviour is connected to a very restricted number of word-forms leads to the following decision. It is necessary, in the computational lexicon, to allow access from the Syntactic and Semantic components beyond the Lemma level to its Inflected word-forms or, better, to the Morphological component of the lexicon.

5. Syntactic properties and correlation between Semantics and Syntax

The first observation made at the syntactic level was that not all the potentially allowable syntactic patterns actually occurred with each inflected word-form. There was a very large range of variation, partially due - as expected - to a difference in the frequency of the word-forms. The most frequent forms obviously present a wider range of patterns. For example, the word-form *dire* (the Infinitive inflection), which is the most frequent, presents almost all the theoretically acceptable patterns. From an analysis of other word-forms, the discrepancy, between what is theoretically acceptable by introspection (i.e. in general the same set of patterns occurring with *dire*) and what is actually found in usage, becomes rather important. We feel that this type of information, which is not of the absolute (yes/no) type, but is more of a preferential nature, is of practical usefulness in a Computational Lexicon and should be recorded with the different inflected word-forms.

This type of annotation (which should in principle be encoded for different text types, since also differences in text types may introduce substantial differences at this level) will become feasible on a large-scale, however, only when automatic tools for syntactic annotation of large Corpora become available. Only such robust tools would make the relevant frequency of the data more easily extractable.

Even though there is a large variation between potentially allowable and actually used syntactic patterns, we did not observe many variations among different word-forms with respect to the frame-element groupings, i.e. variations in the semantic patterns, and in the type of frame elements associated with the inflected word-forms of the verb *dire*.

It is particularly important, for the development of future automatic annotation tools which go beyond surface syntax, i) to establish the relevant correlations between the different surface syntactic elements and the frame elements at the semantic level, and ii) to see how regular these correlations are.

The following correlations always hold:

- Subject corresponds to the Sender/Speaker
- Indirect Object corresponds to Receiver/Addressee
- Direct Object corresponds to Message (of different types)
- di-Infinitive corresponds to Message
- che-Clause corresponds to Message
- wh-Clause corresponds to Message
- direct-quotation corresponds to Message
- PP(su) corresponds to Topic
- PP(di) corresponds to Topic

Despite these regularities, the well-known absence of a univocal mapping between surface syntax and meaning still holds. This is an important drawback with respect to the feasibility of automatically recognising and extracting syntactic patterns and semantic verbal arguments in an unambiguous way. We notice, for example, that the two main meanings of the Italian verb *dire*, i.e. i) "to say something in order to inform about it, to let it know", and ii) "to say something in order to request that something is done", are conveyed by exactly the same surface type of syntactic pattern. These two meanings are exemplified by the following two sentences:

Mark e' sincero quando ***dice a*** *Rachel* ***di*** *amarla.*
(lit: Mark is sincere when he says to Rachel to love her)

Lei ***dice a*** *me* ***di*** *non alzare la voce?*
(lit: You say to me not to raise my voice?)

The di-Infinitive can be interpreted as a Message-type complement in both cases, but the illocutionary force is different, 'declarative' in the first sentence and imperative' in the second.

The first message-type is in fact in free variation with a *che*-clause (that-clause) with Indicative Mood, i.e. we could just as well have:

Mark e' sincero quando ***dice a*** *Rachel* ***che*** *la ama.*
(lit: Mark is sincere when he says to Rachel that he loves her)

We must notice - obviously - that the subject of the Infinitive in the first sentence is the subject of the main clause (subject control), while in the second it is the indirect object of the main clause (indirect object control). This information has therefore to be annotated in a corpus if we want to distinguish between the two main meaning types, even though we see no way of automatically achieving this through use of a parser on a purely syntactic basis.

6. Corpus evidence vs. introspection as a basis for lexicographic description: actual vs. potential

One of the most interesting aspects of using a corpus for a lexicographic task is that one is immediately confronted with the impossibility, based on textual evidence, of using any type of description which is based on a clear-cut boundary between what is admitted and what is not. It is evident that, in the actual usage of a language, a large number of properties are displayed which behave as a *continuum*, and not as properties of a "yes/no" type. In fact, this is one of the main characteristics encountered in actual language usage.

The same holds true for the so-called "rules": we find in corpus evidence more of a "tendency" towards a rule rather than a precise rule. A conclusion which must be drawn from these observations is that almost all information types must not be treated as absolute constraints, whose violation makes a sentence totally unacceptable, but as preferences that make a given sentence more or less acceptable in a given context, without affecting its grammaticality.

This poses a problem at the level of language representation where this type of preferential information has to be accommodated: this may not be easy and certainly not straightforward for a constraint-based formalism. Unfortunately, despite some proposals in this direction, unification (or constraint)-based formalisms, as they exist today, do not easily capture this distinction, i.e. preferences are either ignored or treated as absolute constraints, or else ad-hoc mechanisms are used for dealing with them.

Moreover, the evidence of actual usage is often in contrast with what one would expect if judgement were based solely on introspection. A clear example of this is given by the behaviour of two Italian Speech Act verbs, *chiedere* and *domandare*, both possible translations of the English verb "to ask". They present therefore a problem of lexical selection at the translation level.

These verbs display strong similarities. They are intuitively judged by native speakers to be synonymous and to display quite similar behaviour, also in their polysemy. They are also described in a similar way in traditional dictionaries (where one is frequently used to explain and define the other) and grammars. However, surprisingly, they behave quite differently in the Corpus, producing some unexpected results.

Whereas, theoretically, both of them admit exactly the same type of surface syntactic patterns and related semantic nuances of meaning, they actually display a completely different usage with respect to a number of semantic/syntactic patterns.

While both *domandare* and *chiedere* can convey both the "interrogative (ask to know)" and "imperative (ask to have)" meaning, which seem both quite natural to the speaker, the corpus analysis reveals that - when used with a clausal complement - *domandare* is almost always used in the interrogative sense, and *chiedere* very often in the imperative one. When the imperative message is expressed by a *che-clause,* the verb used is **almost always** *chiedere* throughout the entire corpus (it appears 165 times), although again the same construction with *domandare* cannot be considered as ungrammatical from a theoretical point of view (in fact it is found 3 times).

An even more striking difference between the two verbs is evidenced by the analysis of a homogeneous portion of the corpus, constituted by about 900,000 occurrences from "Il

Sole 24 Ore" (an Italian financial newspaper).The first fact to be noticed is that *chiedere* is used much more than *domandare* (342 occurrences of *chiedere* vs. 35 only of *domandare*), with a much larger disproportion than in the entire corpus (where *domandare* is about one/half than *chiedere*). Moreover, the difference in meaning is even clearer. If we consider the two verbs followed by clausal complements, we find that:

- *chiedere* is used much more in the "request" meaning (82 occurrences, of which 62 in the construction *di* + infinitive, and 20 with *che*-clause) than in the interrogative sense (18 occurrences of wh-clause, of which 12 with *se*, Italian 'whether').
- *domandare* is ***only*** used in the interrogative sense (well 16 occurrences, of which 7 introduced by *se*).

This is a striking example, but many other cases of unexpected - by introspection -evidence can be found.

Again, a (computational or traditional) lexicon has to faithfully represent these facts and these divergences of usage from what is potentially acceptable. Two rules should be followed:

1) The first rule is not to judge what is described in the lexicon on the basis of the native speaker's intuition only, since this leads to a description of a "theoretical language" instead of to a description of the language as it is used.

2) The second rule is to allow, in the lexicon, for a clear representation of (and separation between) what is allowed but only very rarely instantiated, and what is both allowed and actually used.

7. Conclusion

Which are those aspects which should be captured, allowing us to draw lexicographically useful generalisations? A list of aspects were agreed on at different layers of linguistic description (morphologic, syntactic, semantic). When annotated according to these aspects, interesting correlations between different levels, which have to be encoded in a lexicon, emerged.

The analysis of semantic classes made it possible to enucleate the common core of linguistic behaviour which is associated with the broad class, and to differentiate from this the verb-specific aspects related to individual verb-types, or to subclasses of verbs. Particular stress was given to those aspects which were felt to play a critical role in the disambiguation task.

The last step in the project was to go beyond the observational level, in an attempt to generalise, starting from the evidence collected and using it as a basis for a formal modelling of lexical entries.

Only the use of a clear and well-defined methodology, and a classificatory approach which would establish classes of objects with similar behaviours, and which would describe individual classes, allowed the characterisation of the verbs according to a common schema, which in turn could be translated in the formal encoding required by a NLP lexicon, for example the Typed Feature (TF) Representation language chosen in DELIS.

Representation within the TF lexicon highlighted, however, the basic issues and problems encountered when considering the most relevant phenomena resulting from the analysis of the corpus evidence: i) how to determine the appropriate level of abstraction within the type hierarchy for each information type, ii) how to define and represent all the possible interactions between different kinds of information, at different levels of linguistic description, iii) how to encode information types that current versions of HPSG do not easily deal with, such as complex lexical semantic information, collocations, preferences, prototypicality, statistical information, typical modifiers which combine only with particular meanings of a verb, etc. Most of these aspects and phenomena - as seen above -must not be considered on a discrete basis, but rather on a continuum. In order to adequately represent lexical information, a lexical formalism should therefore be able to represent this rather fluid state of affairs.

References

Alonge A., Calzolari N., Monachini M. and Roventini A. (1993), "WP2 - Italian Linguistic Annotation", *DELIS Report*, Pisa.

Bindi R., Monachini M. and Orsolini P. (1991), "Italian Reference Corpus - Key for Consultation", *NERC Working Paper*, Pisa.

Corazzari O., Monachini M., Roventini A. and Calzolari, N. (1996), "Speech Act and Perception Verbs: Generalizations and Contrastive Aspects", in M Gellerstam, J. Järborg, S. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (eds.), Euralex '96 Proceedings, Gothenburg, 73-83.

Federici S. (1993), " A Tool for interactive Frame Assignment ", *DELIS Deliverable*, D-V-1, Pisa.

Fillmore C.J. and Atkins B.T. (1992), "Towards a Frame-based Lexicon: the Semantics of RISK and its Neighbors", in A. Lehrer and E. Kittay (eds.), *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*, Hillsdale, N.J., Erlbaum Associates, 75-102.

Krueger K. and Heid U. (1993), "On the DELIS Corpus Evidence Encoding Schema (CEES)", *DELIS Deliverable*, Stuttgart.

Heid U. (1994), "Relating Lexicon and Corpus Computational Support for Corpus-Based Lexicon Building in DELIS", in W. Martin et al. (eds.), *EURALEX Proceedings*, Amsterdam, 459-471.

Monachini M., Roventini A., Alonge A., Calzolari N. and Corazzari O. (1994),"Linguistic Analysis of Italian Perception and Speech Act Verbs", *DELIS Working Paper*, Pisa.