

Speech Act and Perception Verbs: Generalizations and Contrastive Aspects

Abstract

The present paper fits into the framework of methodologies for corpus-based dictionary building, describing some of the results acquired within the LRE-DELIS Project from the application of an explicitly defined standardized procedure of corpus exploration. Emphasis is given to those aspects which seem to play a crucial role in mono- and multi-lingual lexicography, since they are linked to sense disambiguation. In particular, the general characteristics associated with the two broad classes of perception and speech act verbs and the specific properties related to individual verbs are dealt with, together with considerations on their representation in a hierarchically structured formal language. The relevance of phenomena such as the interaction between different levels of information (morphosyntactic, syntactic, semantic), coindexation, symmetry, typical modifiers, are discussed not only from a mono-lingual point of view, but also from a multi-lingual and contrastive perspective, as a contribution to the selection of the translational equivalent.

1. Introduction

The present paper presents some of the results of the work carried out in the framework of the LRE-DELIS European project (Heid 1994) whose main purpose was to define and experiment a corpus-based lexicographic methodology, in a multilingual environment, for the definition of lexical entries for either NLP or human oriented machine readable dictionaries.

The main purpose of this paper is to show the lexicographic relevance, at both the mono-lingual and multi-lingual/translational levels, of corpus-derived information which is usually neglected or only implicitly expressed by traditional dictionaries. Furthermore, this kind of data may raise interesting problems while encoding it for computational lexicons through a formal hierarchically structured system of features, such as in the Typed Feature Structure (TFS) representation language used in the DELIS framework.

This paper focuses on a subset of Italian verbs belonging to the semantic classes of perception and speech act verbs which have been extensively analysed (Monachini *et al.* 1994) on the basis of the annotated occurrences of the Italian Reference Corpus (Bindi *et al.* 1991).

2. Efficacy of the DELIS Methodology for Lexicography

In the process of building a corpus-based dictionary, the linguist is faced with the problem of (i) defining of the data/aspects to be extracted from the corpus which allow him/her to draw lexicographically useful generalizations and distinctions, and (ii) choosing the methodology suitable for capturing them.

Within DELIS, a list of aspects to be encoded for each verb and its surrounding context was agreed on for all the different linguistic layers: morphosyntactic data – base form and agreement features –; syntactic information – described in terms of functional syntax (subject, object ...) and of surface phrasal structures (NP, PP ...) –; semantic information – encoded from two points of view, selection restrictions on verb arguments (human, concrete ...) and thematic roles according to the theoretical approach of Fillmore's Frame Semantics (Fillmore and Atkins 1992). A Frame Assignment Tool – FAT – (Federici 1993) was implemented for the semi-automatic codification of that information which could not be automatically encoded.

The methodology for corpus annotation agreed on by all the partners is outlined in the CEES - Corpus Evidence Encoding Schema – (Krueger and Heid 1993). This schema allows us to (i) gather in a unique frame – in an HPSG-like approach – the properties belonging to different linguistic levels of the item under analysis, enabling us to computationally retrieve them singularly or in correlation; (ii) encode these properties according to a standardized procedure, facilitating the comparison of the information in a multi-lingual environment.

The encoding of the semantic deep level permits us to collect together, under the same label, different surface syntactic realizations having the same role (e.g. the Percept in Italian which can be syntactically expressed by an NP or a VP). Hence, generalizations, correlations and comparisons can be established with obvious advantages not only from a mono-lingual perspective, but also from the point of view of the multi-lingual and contrastive lexicography.

The DELIS approach made it possible to enucleate the common core of the linguistic behaviour associated with a broad semantic class – e.g. the perception class is characterized by the Experiencer and the Percept roles, while the speech act class displays the three roles of Sender, Message and Receiver – and, at the same time, to discover the specific properties related to individual verb types – e.g. a few verbs in the perception class, depending on both morphological and syntactic features, shift towards the speech act class. In this way, particular stress was given

to those aspects which seem to play a critical role in lexicography, since they are linked to sense disambiguation.

This methodology, finally, allowed us to characterize these verbs according to common properties: this constitutes the first step towards translation into a formal encoding required by a NLP lexicon and, in particular, into the TFS representation language agreed on in DELIS.

3. Speech Act and Perception Verbs

The analysed subset of Italian speech act verbs includes *dire* (to say, to tell), *promettere* (to promise), *domandare* (to ask), *chiedere* (to ask), *discutere* (to discuss) and *negare* (to deny). These verbs were chosen because they seemed to cover a sufficient range of typical phenomena of the communication domain. *Dire*, in particular, was chosen because it is the more general verb in the field, somehow representing the field itself. The other ones were chosen because they are more specific with respect to *dire* (to say) and display additional or different properties (Heid *et al.* 1993).

Speech act verbs present, by their very nature, a different semantic and syntactic behaviour with respect to perception verbs. Firstly, from the semantic point of view, they do not present a high degree of polysemy (if we do not consider their appearance in multiwords or in idioms). Shifts of meaning, sometimes determined by the type of direct object, do not usually change the nature of the speech act verb, but give rise to synonymous verbs (e.g.: *dire preghiere* = *pregare*; *dire ciao* = *salutare*; *dire bugie* = *mentire*; *promettere guerra* = *minacciare*; *chiedere scusa* = *scusarsi* etc.). Secondly, from the syntactic point of view, speech act verbs occur in a rather large variety of subcategorization patterns, but very few of these are a clear sign 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 subset of analysed perception verbs includes *sentire* (to feel), *vedere* (to see), *udire* (to hear), *ascoltare* (to listen), *guardare* (to look), *gustare* (to taste) and *annusare* (to smell). *Sentire*, as with *dire* for speech act verbs, is the more general and representative verb within the perception class because, by selecting various types of internal arguments, it can be used to express all the modalities except vision (e.g. *sentire il sapore*, to taste; *sentire il suono*, to hear; ...). In contrast to speech act verbs, perception verbs do not appear very homogeneous from a syntactic point of view and differ greatly in frequency, degree of

polysemy and internal argument realization. Regarding their thematic structure, they are characterized by an Experiencer role which can be active or passive and by a Percept role which includes various subtypes (Source, Stimulus, Inference, Event, etc.) corresponding to various syntactic realizations. On the basis of semantic features, a first raw subdivision can be made within the class: the type of Experiencer distinguishes *vedere*, *sentire* and *udire* (passive) from *ascoltare*, *guardare*, *gustare* and *annusare* (active).

When compared with other members of the class, high frequency and richness of syntactic complementation, linked to a considerable number of word-senses, characterize *sentire* and *vedere*. In a few cases these two verbs, even if they represent the passive way of perceiving, can acquire an active value, depending on morphological and syntactic/aspectual features, shifting towards the corresponding active verbs, with implications at the translational level. They also present shifts of meaning from the perception class to the speech and knowledge verbs, which appear to be determined by a different sentential complementation: *senti se il treno e' partito = chiedere* (to ask), *senti' il treno partire = udire* (to hear), *senti' che il treno era partito = venire a sapere, apprendere* (to learn).

Gustare and *annusare* are the most specific verbs and present a simple argument structure; their only noticeable characteristic is their frequent figurative use (also attested in some deverbal perception nouns, as discussed in Roventini and Monachini 1995).

4. Examples of Generalizations and Contrastive Problems

4.1 Corpus-derived information about meaning distinction

In analyzing the annotated corpus, one of the most interesting aspects we noted was that, to capture all the nuances of the meanings of a word and its syntactic-semantic behaviour, one must go beyond its representation in terms of phrasal types, grammatical functions and semantic roles because, in most cases, many other data types contribute to meaning distinctions. These, unlike the previous ones, are more difficult to capture as, for example:

- the combination and the interaction of morphological, morpho-syntactic, syntactic and semantic data;
- mere syntactic data such as coindexation, symmetry etc.;

- specific syntactic-semantic information not referring to arguments, such as the typical modifiers, etc.

The relevance of these data emerges at different levels: (i) at the multilingual level since, by highlighting specific idiosyncratic meanings of the verb, this kind of information is crucial for selecting the correct translational equivalent and (ii) at the monolingual level since, by incorporating and formalizing these data into a NLP system, they contribute to the semantic analysis of the text and to the enucleation of the deep semantic content.

Interesting examples of relevant interactions between different levels of information are those cases in which a verb acquires, with specific inflected wordforms, a very specific meaning, not found with other inflections. For instance, *direi*, 1st Singular Present Conditional (I would say), which in many occurrences is used in a parenthetical construction, acquires the meaning of the adverbial multiword expression *per così dire* (as it were). E.g.:

<i>lo sguardo era sincero, direi amichevole</i>	(his glance was sincere, I would say friendly)
<i>in modo direi scorretto</i>	(in a way, I would say, incorrect)

Sometimes meaning distinction is given by particular inflections only if they are combined with specific syntactic data, such as the type of syntactic pattern in which the verb occurs. This is the case with *diresti* (2nd singular Conditional Present) ‘you would say’ which has the meaning of *ti piacerebbe?* ‘would you like?’ when it occurs in the construction ‘*Cosa/che ne diresti + di-Infinitive(subj. control) / + PP-di?*’. E.g.:

<i>Cosa ne diresti di andare a mangiare qualcosa?</i>	(What would you say to go out and eat something?)
<i>Cosa ne diresti di un panino?</i>	(What would you say about a sandwich)

At the translational level this expression has an exact French equivalent, but, as far as English is concerned, it should be translated into syntactically different sequences such as: ‘what do you think of ...?/what about ...?’.

In the perception domain, interesting examples of the same phenomenon are the 2nd sing/plur of the Imperative and the 3rd sing of the Subjunctive (polite form) of *sentire*, *ascoltare*, *vedere* and *guardare* in the so-called “allocutive” construction used, in a colloquial register, to

establish contact between speakers or to recall the attention of the listener to something.

<i>Senta, prenda questo, per favore</i>	(Listen, take this, please)
<i>Guarda: se fossi una ragazza onesta</i>	(Look: if I were an honest girl)
<i>Vedi, dice, devi solo avere pazienza</i>	(You see, he says, you only have to have)
<i>Ascolta, dobbiamo finire la casa</i>	(Listen, we have to finish building the house)

It is worth noting that, in these constructions, even if the equivalent French/English perception verbs constitute an admitted literal translation, sometimes the correct equivalent is difficult to find, or there may be doubts about the “best” translation: indeed, the best translation may be a completely different construction. In the first of the above examples, *sentire* is not translated by the French verb ‘sentir’ but by ‘écouter’. The same holds for English, where ‘to listen’ is the correct translation. In the second example, *guardare*, in an allocutive construction, cannot be translated by the French words ‘regarde/regardez’, but only by ‘écoute/écoutez’.

Also the verb *guardare*, used in the Imperative form followed by a *che*-clause, does not indicate the visual modality and cannot be translated by the French verb ‘regarder’, but can be rendered by ‘écoute’ or, depending on the linguistic register, also by the interjection ‘attention ...!’. As far as English is concerned, native speakers, asked to provide a translation, gave us the following different interpretations:

<i>guarda che non hai capito</i>	(look, you have not understood) – 1st transl. – (I’m sorry you haven’t understood) – 2nd transl. –
----------------------------------	---

This proves that there are contrastive problems also due to the syntactic peculiarities of this construction.

As we have already mentioned, also the presence of a symmetrical construction and coindexation can be of help in resolving meaning distinction. For instance, the perception verbs *vedere* and *sentire*, used in symmetrical constructions of the type

- (i) *X and Y si vedono*
- (ii) *X and Y si sentono*

have the meaning, respectively, of *incontrarsi* ‘to meet’ and *parlarsi* ‘to speak on the phone’. The French equivalent ‘voir’, as well as the English ‘to see’, having also the sense of ‘rencontrer’ and ‘to meet’, constitute admitted translations for the expression of type (i); whereas the French

'sentir' and the English 'to hear' do not convey the same meaning and therefore another verb should be found for translating the expression of type (ii).

On the other hand, in the communication domain, coindexation between the Subject of the main clause and that of the subordinate clause is particularly relevant e.g. to distinguish those cases in which verbs such as *dire/domandare* 'to say/to ask' acquire the meaning of 'to order'. Indeed, the construction *di*-Infinitive carries an Imperative Message only if its Subject is different from that of the main clause (example b.):

- a. *dice di essere stanco* (declarative)
(he says he is tired)
- b. *dice di allontanarci immediatamente* (imperative)
(he is telling us to move back immediately)

The relevance of this information from a translational point of view is questionable since the ambiguity is sometimes preserved in other languages (see English and French). It is, however, essential in any type of system designed, with a view towards the future, for the semantic analysis and interpretation of a text.

To give a further example, the identification of particular meanings of the verb can also rely, as we have already mentioned, on specific semantic information such as the type of modifier combined with the verb. Indeed, in some cases, different Manner modifiers seem to select different meanings of the verb. This is the case with *discutere*, which means (i) 'to argue' or (ii) 'to treat/analyze' when combined with the following modifiers:

- | | |
|---|--|
| <i>discutono accanitamente</i> | (they are arguing bitterly) – 1st sense – |
| <i>discutono in dettaglio il problema</i> | (they are analyzing the problem in detail) |
| | – 2nd sense – |

In the perception class, different types of adjuncts can help in the recognition of the active and the passive perception: (i) manner adjuncts which focus on the mental attitude and participation of the Experiencer are typical of the active perception verbs; (ii) time adjuncts which express the immediacy of the perception are linked to the passive ones. E.g:

- | | |
|---|---|
| <i>stava a sentire attentamente</i> | (he was listening with attention) – active – |
| <i>in quel momento senti' il rumore</i> | (in that moment he heard the noise) – passive – |

It is worth noting that, in the particular case above, where one verb in the source language presents two correspondents in the target language, such as English 'to listen' and 'to hear' (corresponding to the active and passive modality respectively) and French 'écouter' and 'entendre', the presence of adjuncts suggests the correct translation.

4.2 Corpus-derived data for selecting between synonymous

Up to now we have focused on corpus-derived data which are crucial for the identification of particular idiosyncratic meanings of the verb. However, other relevant information can be obtained from the analysis of a considerable number of occurrences.

Interesting examples are the statistical and frequency data which enable one to distinguish between two lexical items usually given by traditional lexicography, paper dictionaries and even by native speakers' intuition as synonymous and perfectly interchangeable. For example, *chiedere* 'to ask' instead of *domandare* 'to ask' seems to be strongly preferred in particular syntactic constructions and in particular lexical combinations. A comparison of the frequency data of the two verbs with a given set of constructions and lexical items allow us to distinguish them on a more sound basis. From a multilingual perspective, this information is of utmost importance for the selection of the correct translation, for instance when translating the English 'to ask' or the French 'demander' into the Italian *chiedere/domandare*.

Chiedere sounds more natural, from an analysis of actual usage, than *domandare* for conveying a Message with Imperative illocutionary force:

Lui le chiede di aspettarlo (He asks her to wait for him)

However, the verb *domandare* may also be used, although much less frequently, with the syntactic construction *di*-Infinitive.

On the other hand, if the Imperative Message is expressed by a syntactic construction of the type *che*-Clause, the verb which is uncontroversially preferred for introducing it, throughout the entire corpus, is *chiedere*, although the same construction with *domandare* cannot be considered ungrammatical from a theoretical point of view.

La Fipe chiede che il periodo di ferie sia portato a 15 giorni

(Fipe asks that the vacation period be brought to 15 days)

Dubcek chiede che gli venga restituito il suo onore politico

(Dubcek asks that his political honour be given back)

Messages with Interrogative illocutionary force (indirect questions) are most frequently introduced by *domandare*. For example:

<i>si domanda chi sara' il prossimo presidente</i>	(He is wondering who will be the next president)
<i>gli domandano come e' stato reclutato</i>	(They are asking how he has been recruited)

Further differences can be identified at the level of lexical co-occurrences. The two verbs are totally interchangeable in many word combinations, but some lexical items select only one of them. Indeed, *chiedere* can always replace *domandare*, but the reverse is not always possible.

*chiedere/*domandare la rivincita/ un colloquio/ grazie/ la linea ...*
 (ask for a return game/an interview/thanks/a telephone line)

A parallel, but reversed, example is that of *sentire*, in the perception domain of audition, which covers both the English 'to hear' and 'to listen' (Monachini and Roventini 1994). Those contexts in which *sentire* corresponds to *ascoltare*, (i.e. the active modality) and therefore should be translated with 'to listen', present the allocutive/exhortative construction (see above) or the infinitive form in dependence of motion and aspectual verbs:

<i>andare a sentire</i>	(to go to listen)
<i>mettersi a sentire</i>	(to start listening)

At the contrastive level, those cases in which *sentire*, followed by the *che*-clause, acquires the cognition meaning of *venire a sapere* 'to know' do not cause problems for the English equivalent, since 'to hear' presents the same shift. The same, however, is not true for French, where for this sense 'entendre' is used (and not 'ecouter').

sentì che la casa era venduta (he heard that the house was sold)

5. Conclusive Remarks

The DELIS corpus-based methodology proved to be of the utmost importance and usefulness in a lexicographic environment. This paper does not cover all possible corpus-based relevant information, but it gives, nevertheless, a general view of just many linguistic aspects can

determine – individually or in combination – the lexical meaning of words, and why efforts should be devoted to their extraction and formalization.

In particular, the discussion in this paper focused on some of the information derived from the analysis of the annotated corpus which are relevant to meaning distinction. The most interesting observation, derived from the data emerging from an analysis of the corpus, is that meaning distinction cannot always rely on the information taken from phrasal types, grammatical functions and their thematic roles. Many idiosyncratic meanings can be enucleated only by taking other information into account, usually missing in traditional paper dictionaries, at the level of morphosyntax, semantics, statistics and the interactions between different levels of information. Such data becomes particularly relevant from the perspective of a possible future fully automated lexical knowledge extraction process.

Another of the most interesting aspects of the use of a corpus for lexicographic purposes is the immediate evidence of the impossibility to use any type of description which is based on a clear-cut boundary between what is admitted and what is not. In actual usage of the language it is evident that its main characteristic is that of displaying a large number of properties which behave as a continuum, and not as properties of 'yes/no' type. The same can be said for the so-called "rules", where we find in corpus evidence more a tendency towards a rule than a precise rule. All this type of information must not be treated as absolute constraints, whose violation makes a sentence totally unacceptable, but rather as preferences, that make a given sentence more or less acceptable in a given context without affecting its grammaticality (Calzolari, forthcoming).

Future developments of a corpus-based lexicographic approach, therefore, should aim to enucleate this kind of information and express it in formal terms within a computational lexicon: the representation language, i.e., should be able to accommodate this type of preferential information and clearly represent and separate what is allowed but only very rarely instantiated, and what is allowed and actually used.

References

- Bindi, R., M. Monachini, P. Orsolini 1991, "Italian Reference Corpus – Key for Consultation", NERC Working Paper, Pisa.
- Calzolari, N. (forthcoming), "Observation and Generalization: Corpus-based Linguistic Analysis of Italian Speech Act Verbs", Pisa.
- Federici, S. 1993, "A Tool for Interactive Frame Assignment", DELIS Deliverable, D-V-1 Pisa.
- Fillmore, C.J. and B.T. Atkins 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, pp. 75–102.
- Krueger, K. and U. Heid 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 *EURALEX94 Proceedings*, W. Martin et al. (eds.), Amsterdam, pp. 459–471.
- Heid, U. et al. 1993, "Towards work on speech act verbs in DELIS", Delis Internal Report, Stuttgart.
- Monachini, M. and A. Roventini 1994, "Italian Audition Verbs: a Corpus- and Frame-based Analysis", in *EURALEX94 Proceedings*, W. Martin et al. (eds.), Amsterdam, pp. 117–127.
- Monachini, M., A. Roventini, A. Alonge, N. Calzolari and O. Corazzari 1994, "Linguistic Analysis of Italian Perception and Speech Act Verbs", in Ostler N. (ed.), DELIS-Deliverable, D-II, Pisa and London.
- Roventini, A. and M. Monachini 1995, "Perception Nouns in the Italian Reference Corpus: Argument Structures and Collocational uses", in *ACH-ALLC95 Proceedings*, University of California, Santa Barbara, July 11–15, pp.101–105.