Short summary of the ELSNET Syntax/Semantics Annotation Task

Ornella Corazzari, Nicoletta Calzolari, Antonio Zampolli (Pisa)
and
Ulrich Heid and Hannah Kermes (Stuttgart)

Overview

In 1998, ELSNET started a small-scale resource-building exercise, to produce a mini-corpus of sentences annotated at both syntactic and semantic level. The resource produced is a pair of small sample corpora of parallel structure for German and Italian, about 1000 sentences of each language, illustrating 20 verbs, and their syntactic and semantic subcategorization. The annotation concentrates on the verbal predicates and their subcategorized complements, as well as on a few relevant modifiers.

This resource is to be seen, among others, as an example of a corpus annotated at several levels of linguistic description, including word classes, syntax and lexical semantics. As it is the case with all experimental resources produced by the ELSNET Resources Task Group, the corpora are intended to be an example of the kind of resources that may, in the medium term, be produced in a wider framework. Thus, the production methodology is seen as equally important as the data outcome.

Objectives

The following are the main objectives of the annotation task:

- Increasing the knowledge about semantic corpus annotation, in particular with a view to evaluate the usability of the classifications embodied in semantic descriptions of recent European projects, such as EuroWordNet and SIMPLE. Even though for German, other than for Italian, no full taxonomy is available, an attempt was made to use the same inventory of semantic types (derived from SIMPLE), for both languages, and this task has indeed proven feasible without greater obstacles.
- Analyzing to what extent syntactic and semantic information can be used for word sense disambiguation, and, in the long run, for semi-automatic alignment of equivalent candidates. Therefore, a tabular summary of the data will be generated (see database dump).
- Determining to what extent a corpus annotated with both syntactic and semantic information serves purposes of a detailed lexicographic description of the verbs analyzed. With a view to work on collocations, semantic selection restrictions, as well as constraints for use in applications of NLP (e.g. in the fields of Information Extraction and Information Retrieval), it is useful to have a small scale resource available, which can serve as a test bed for different approaches in this area.

- Testing available tools for query and retrieval of data from corpora with multi-layered annotation: the data output of the ELSNET action is a corpus with multi-layered annotation, which can be used as a test resource for query tools, such as the MATE Workbench (designed in the MATE project, which itself is the result of an ELSNET exploratory action). A version of the data compatible with the MATE guidelines for linking of elements is under construction.

The corpus is experimental in several respects, in particular because of its small size, and of the specific conditions on item selection.

Corpus characteristics

A comparable Italian/German corpus was created, in so far as both, corpus composition and annotation, are based on the same principles for both languages:

- Around 20 verbs of Italian have been selected, mainly to cover different semantic fields and a variety of syntactic constructions, and to have examples of a medium degree of polysemy. On that basis, around 20 German verbs were selected, which are known to be (partial) translation equivalents of the Italian verbs. Again, emphasis was on a medium degree of polysemy, and on rich documentation in the source corpora. In five cases, two German verbs have been selected which are partial equivalents of one single Italian item, to facilitate experiments in translation equivalent selection and word sense disambiguation. A table summarizing the selected verbs is appended, for both languages (cf. appendix).
- The corpora serving as raw material are mainly journalistic: for Italian, the following local or national sources have been used: *Il Corriere della Sera*, *La Repubblica*, *La Stampa*, *Il Sole 24 Ore*, *Unione Sarda* and *Panorama*. For German, *Frankfurter Rundschau*, *die Tageszeitung*, and *Stuttgarter Zeitung* are the main sources: again local or national newspapers.
- For each Italian verb around 50 sentences have been selected from the data. For each German verb, it was aimed at the same number; in a few cases, somewhat less material was available, which was only accepted when two German verbs shared an Italian equivalent.
- For semantic annotation, basically verbs and their arguments noun groups (i.e.) and prepositional groups are of interest. Consequently, emphasis was laid on non-sentential readings, and readings with sentential complements were only used as auxiliary information, to complete the syntactic description, wherever necessary.

Annotation: Schemata used and procedures applied

The annotation available in the corpus covers morphosyntax, grammatical functions, as far as they are relevant for the description of the verbal predicates, as well as semantic features attached to heads of relevant phrases.

For morphosyntax, the annotation follows the EAGLES/PAROLE guidelines. For the syntactic layer, the SPARKLE/MATE guidelines have been followed. For the semantic annotation, it was decided, at least for the Italian part of the corpus, to have a double annotation:

1. a "word sense" annotation, through the linking of each corpus occurrence of a verb EuroWordNet readings;
2. a "semantic" annotation, through the assignment to each corpus occurrence of acomplement noun to the semantic types of the SIMPLE antology.

As for German, WordNet data on verbs were not available at the time of compilation of the corpus, only the second (antological) step was performed.

For German, the following steps were performed automatically: the selection of relevant sentences, part of speech tagging, annotation of grammatical functions: a stochastic part of speech tagger was used ([Schmid 1994]) and thereafter, the resulting data were processed by means of an LFG-based grammar of German, which assigns grammatical function labels. The grammatical function annotation from LFG was mapped onto the SPARKLE/MATE standards proposal, and the resulting MATE-conformant annotation was converted to XML.

A number of scripts are available which were used to perform the individual annotation steps, such that more material can in principle be processed according to the same procedures, whenever this becomes necessary.

Evidently, a completely manual correction step was necessary after syntactic analysis (elimination of wrong analyses, elimination of useless ambiguities), and, moreover, semantic annotation was carried out completely manually. This has in particular to do with the lack of semantic resources for German.

For Italian, part-of-speech-tagging was performed automatically.

This was followed by the marking of grammatical relations and the assignment and manual selection of the relevant EuroWordNet and simple tags (for wordsense and semantic type, respectively).

Form of the delivery

The data are delivered in the form of an annotated corpus, encoded in XML, according to MATE guidelines. It is however planned, to have a few analyses of the available material follow the actual data production, such that more insight can be gained on the basis of the resource. However, to be sure that the resource is accessible to interested parties at present, already, the current "undigested" version is made available via ELRA (under discussion, as of May 2000).

A first type of analysis will consist in the provision of tabular output, indicating for each verb all possible syntactic subcategorization types, along with an inventory, for each subcategorization type, of the semantic types found in the nominal groups that realize the

functions. This will be a preliminary version of a corpus-documented syntactic and semantic dictionary, as it could be the outcome of a large scale semi-automatic annotation exercise, based on taxonomic resources of the kind of SIMPLE or EWN.

This small collection of entries will, we hope, provide insight into the new lexical resources which could be extracted from corpora of the kind illustrated here.

List of Italian and German verbs described

Note: In the German part, "SC:n" indicates the number of different subcategorzation frames identified. If "reflexives" were found, these are noted separately.

List of selected verbs	
Italian	German
abbandonare	aufgeben SC:4[1] ; verlassen SC:2
arrestare	festnehmen SC:1; aufhalten SC:1+3refl
aprire	öffnen SC:2+3refl; eröffnen SC:3+1refl
prevedere	vorsehen/planen SC:2+1(2)refl/SC:2; vermuten/ahnen SC:1/SC:2
tagliare	reduzieren SC:2+1refl; unterbrechen SC:3(4)
comprendere	verstehen SC:3(4)+2(3)refl
chiamare	rufen SC:4(6); aufrufen SC:(3)
chiedere	bitten SC:3; fragen (nach) SC:2+2(nach)+1refl
perseguire	fortsetzen SC:2+2refl; verfolgen SC:2
coprire	umfassen SC:2+2refl; decken; schützen SC:4(6)+2(4)refl
entrare	passen (in) SC:3+1(in)+1(zu); eintreten SC:3(4); sich beteiligen SC:2+2refl
vedere	sehen SC:4(5)+1refl; erkennen SC:3(5)+1refl
mantenere	halten SC:7(8)
mantenersi	andauern SC:2; sich halten (an) SC:2(3)
percepire	erhalten SC:1
esercitare	tätig sein/betreiben SC:2; üben/sich üben SC2+2refl
presentare	zeigen SC:5+2refl; präsentieren SC:3(4)+2refl
portare	bringen SC:4(5); tragen SC:4+1refl
provare	üben/proben SC:3; versuchen SC:1+2(4)refl; spüren/empfinden SC:3/SC:2(3)
realizzare	realisieren SC:1; begreifen SC:3
arrivare	erreichen SC:1; ankommen SC:6(8)