Proc. Second Workshop on Treebanks and Linguistic Theories, 14–15 November 2003, Växjö, Sweden (TLT 2003), pp. 213–216

FuSe – a Multi-Layered Parallel Treebank

Lea Cyrus, Hendrik Feddes, Frank Schumacher

Arbeitsbereich Linguistik, University of Münster
{lea,feddes,frank}@marley.uni-muenster.de

1 Overview

While there exist a number of bi- and even multilingual corpora, syntactically analyzed parallel corpora are rare.[1] At Münster University, we have initiated a treebank project with the aim of closing this gap. Our goal is to build a multi-layered treebank of aligned parallel texts in English and German. While we confine ourselves to annotating only one language pair, the design will be such that additional languages can be added, provided there exist appropriate translations.[2] Our working title for the treebank is FuSe, which stands for *fu*nctional *se*mantic annotation and connotes that two or more languages are *fused* with each other. Although our main motivation is to contribute to linguistic research rather than to develop a corpus which is tailor-made for a particular NLP-application, we believe that the corpus will prove useful for research in several fields of application, the most obvious one being machine translation.

The linguistic annotation of the FuSe corpus will contain the following layers: POS tags, constituent structure, functional relations, predicate-argument structure, and alignment information. The alignment layer is the only one which is defined for a language pair rather than for a single language. Apart from this layer, the subcorpora are complete monolingual resources in their own right. In the following we will concentrate on the predicate-argument structure and on the representation of alignment information.

[1] In the Parallel Grammar Project as described in [1], the term “parallel” is used in the sense that similar phenomena in the languages under investigation are represented in a similar – parallel – way. The only treebank project we know of that understands “parallel” to mean that the texts in the respective languages are translations of each other, is the Korean-English treebank mentioned in [2].

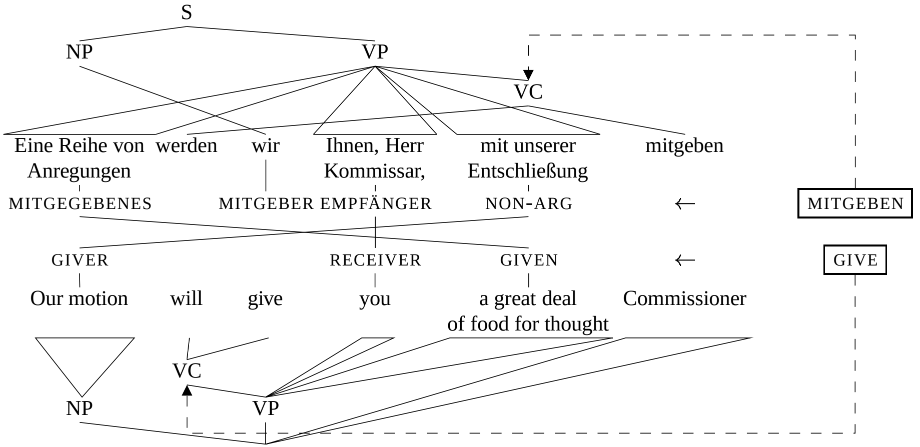
[2] We use Philipp Koehn’s sentence-aligned Europarl corpus [3], which, being available in eleven languages, gives ample opportunity for extending the treebank. In the following, examples from this corpus are identified by filename and line number.

2 Predicate-Argument Structure and Alignment

Alignment of parallel corpora is typically done on the sentence level, where we do not expect a $1:1$, but an $n:m$ relationship between sentences, with rather low numbers for n and m, as in (1), where one German sentence corresponds to two English ones.

(1) a. Unser Parlament hat diesem Text zugestimmt, allerdings mehrere Änderungsvorschläge eingebracht [. . .]. (de-en/de/ep-00-01-18.al, 1034)
b. Parliament voted in favour of this text. It did, however, table a number of amendments [. . .]. (de-en/en/ep-00-01-18.al, 1034)

With parallel treebanks, the possibility opens up to align texts on the level of constituent structure, which is more fine-grained than the sentential level and makes the treebank a much richer resource for many purposes. However, when sentences have a completely different constituent structure, as in the example below, alignment beyond sentence level seems impossible. In order to arrive at a linguistically more interesting alignment, we aim at establishing a basic predicate-argument structure for both sentences which we can then use for bilingual alignment of corresponding predicates and arguments. This can be illustrated as follows:[3]



(de-en/ep-00-01-18.al, 53)

The annotation has to be simple enough to be manageable,[4] yet rich enough to be of use for the alignment. We regard the following expressions as candidates

[3] Irrelevant details are omitted. VC stands for “verb complex”.

[4] Particularly, we do not attempt to represent the sentence as a whole in an interlingua-like fashion. Also, the predicates in a given sentence are not connected with each other. In many cases, nesting of predicates can be derived from the constituent structure, with which these predicates are connected.

for representing predicates: verbs, adjectives, and nouns subcategorizing other elements (e. g. deverbal nouns). Predicates and arguments are recorded in a predicate-argument database. Predicates are represented in the database by the citation form of the corresponding word token in the corpus. Arguments are given short intuitive role names whose only purpose it is to distinguish the arguments of a particular predicate type and to group similar arguments of different predicate tokens which represent the same type. No further attempts at generalization in a FrameNet-like[5] manner are made. If it turns out that an argument in one language corresponds to a constituent that has not been marked as an argument in the other language, this constituent is given a pseudo-tag (NON-ARG) during the alignment process.

We annotate only those arguments which are present in the text. We distinguish four kinds of “presentness” of arguments. The first type of argument is syntactically present and appears in its standard form in the same clause as its predicate. Arguments of this kind are uncontroversial and hence enter the database record directly and unmarked.

The second type of argument is syntactically present, but does not appear in the same clause as its predicate due to syntactic constructions such as raising and control. Arguments of this kind enter the database record for this predicate, but are marked accordingly, so as to account for the objective case of the constituent functioning as subject in the infinitival clause in sentences like (2).

(2) Allow [$_{\text{argument STARTER}}$ me] to [$_{\text{predicate START}}$ start] with the most important demand. (de-en/en/ep-00-01-18.al, 2932)

A further type of argument is only implicitly present, e.g. the “logical” subject in imperatives or passive constructions. We do not mark arguments of this kind, and they do not enter the database. However, we plan to mark passive voice and imperative sentences in the constituent structure so as to be able to filter queries or to automatically enrich predicate-argument lexicons generated from the corpus. Thus, the fact that the predicate-argument structures of sentences like (3) lack an important argument (the DESTROYER) could be traced back to their being derived from sentences in the passive.

(3) Natural habitats were destroyed. (de-en/en/ep-00-01-18.al, 1836)

The last type of argument is only vaguely present in another clause or even sentence. A human annotator could spot the argument, but not necessarily bind it to a constituent. Due to the vagueness involved, we do not mark arguments of this kind at the moment:

[5] See `http://www.icsi.berkeley.edu/~framenet` and the references listed there.

(4) I only hope my remarks will not harm your chances of re-election. Unfortunately the citizens themselves do not always appreciate the good that is being done for them, especially here. (de-en/en/ep-99-02-09.al, 1222f.)

In the first sentence in (4), the nominal predicate *re-election* lacks the argument ELECTOR. It could be argued that it is the *citizens* mentioned in the next sentence who will perform the act of re-electing, but our annotation does not cross sentence-boundaries, hence this argument is not recorded.

During monolingual annotation, the annotators mark all the constituent tokens in a sentence which are candidates for representing predicates, thereby prompting the annotation tool to look up the predicate-argument database to check whether the same[6] predicate has been seen before. If this is the case, the annotator is presented with the argument roles found so far. Deverbal nouns are not merged with their corresponding verbs as belonging to one predicate even where this seems possible. However, the predicate-argument database records links between the two predicates so as to enrich the choice of possible argument roles presented to the annotator during the annotation process. Unseen predicates and argument roles are added to the database by the annotator. While we use Oliver Plaehn's *Annotate*[7] for phrasal and functional annotation and for manual corrections of the POS tags, the predicate-argument database and additional tools for handling the predicate-argument structure and the alignment layer are currently under development.

References

[1] Butt, M., S. Dipper, A. Frank and T. H. King (1999) Writing large-scale parallel grammars for English, French, and German. In M. Butt and T. H. King (ed.) *Prodeedings of the LFG99 Conference*, pp. 1–59. Stanford: CSLI Online Publications.

[2] Dras, M. and C. Han (2002) Korean–English MT and S-TAG. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6).* (URL: `http://www.sfu.ca/~chunghye/papers/tag6-stag-paper.pdf`).

[3] Koehn, P. (2002) Europarl: A Multilingual Corpus for Evaluation of Machine Translation. (URL: `http://www.isi.edu/~koehn/publications/europarl`).

[6] It could be necessary here for the annotator to choose from a set of "homonyms". Homonymous predicates are distinguished by short glosses. As with the argument roles, this is done on the fly by the annotators and has the sole purpose of intuitive differentiation.

[7] See `http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html`.