A Model for Fine-Grained Alignment of Multilingual Texts

Lea CYRUS and Hendrik FEDDES*
Arbeitsbereich Linguistik
University of Münster
Hüfferstraße 27, 48149 Münster, Germany
{lea,feddes}@marley.uni-muenster.de

Abstract

While alignment of texts on the sentential level is often seen as being too coarse, and word alignment as being too fine-grained, bi- or multilingual texts which are aligned on a level in-between are a useful resource for many purposes. Starting from a number of examples of non-literal translations, which tend to make alignment difficult, we describe an alignment model which copes with these cases by explicitly coding them. The model is based on predicate-argument structures and thus covers the middle ground between sentence and word alignment. The model is currently used in a recently initiated project of a parallel English-German treebank (FuSe), which can in principle be extended with additional languages.

1 Introduction

When building parallel linguistic resources, one of the most obvious problems that need be solved is that of alignment. Usually, in sentence- or word-aligned corpora, alignments are unmarked relations between corresponding elements. They are unmarked because the kind of correspondence between two elements is either obvious or beyond classification. E. g., in a sentence-aligned corpus, the $n : m$ relations that hold between sentences express the fact that the propositions contained in n sentences in L1 are basically the same as the propositions in m sentences in L2 (lowest common denominator). No further information about the kind of correspondence could possibly be added on this degree of granularity. On the other hand, in word-aligned corpora, words are usually aligned as being "lexically equivalent" or are not aligned at all.[1] Although there are many shades of "lexical equivalence", these are usually not explicitly categorised. As (Hansen-Schirra and Neumann, 2003) point out, for many research questions neither type of alignment is sufficient, since the most interesting phenomena can be found on a level between these two extremes.

We propose a more finely grained model of alignment which is based on monolingual predicate-argument structures, since we assume that, while translations can be non-literal in a variety of ways, they must be based on similar predicates and arguments for some kind of translational equivalence to be achieved. Furthermore, our model explicitly encodes the ways in which the two versions of a text deviate from each other. (Salkie, 2002) points out that the possibility to investigate what types of non-literal translations occur on a regular basis is one of the major profits that linguists and translation theorists can draw from parallel corpora.

In Section 2, we begin by describing some ways in which translations can deviate from one another. We then describe in detail the alignment model, which is based on a monolingual predicate-argument structure (Section 3). In Section 4 we conclude by introducing the parallel treebank project FuSe which uses the model described in this paper to align German and English texts from the Europarl parallel corpus (Koehn, 2002).

2 Differences in Translations

In most cases, translations are not absolutely literal counterparts of their source texts. In order to avoid translationese, i. e. deviations from the norms of the target language, a skilled translator will apply certain mechanisms, which (Salkie, 2002) calls "inventive translations" and which need to be captured and systematised. The following section will give some examples[2]

[*] We would like to thank our colleague Frank Schumacher for many valuable comments on this paper.

[1] Cf. the approach described in (Melamed, 1998).

[2] As we work with English and German, all examples are taken from these two languages. They are taken from the Europarl corpus (see Section 4) and are abbreviated where necessary. Unfortunately, it is not eas-

of common discrepancies encountered between a source text and its translation.

2.1 Nominalisations

Quite frequently, verbal expressions in L1 are expressed by corresponding nominalisations in L2. This departure from the source text results in a completely different structure of the target sentence, as can be seen in (1) and (2), where the English verb *harmonise* is expressed as *Harmonisierung* in German. The argument of the English verb functioning as the grammatical subject is realised as a postnominal modifier in the German sentence.

(1) The laws against racism must be harmonised.[3]

(2) Die Harmonisierung der
The harmonisation of_the
Rechtsvorschriften gegen den
laws against the
Rassismus ist dringend erforderlich.
racism is urgently necessary.

This case is particularly interesting, because it involves a case of modality. In the English sentence, the verb is modified by the modal auxiliary *must*. In order to express the modality in the German version, a different strategy is applied, namely the use of an adjective with modal meaning (*erforderlich*, 'necessary'). Consequently, there are two predications in the German sentence as opposed to only one predication in the English sentence.

2.2 Voice

A further way in which translations can differ from their source is the choice of active or passive voice. This is exemplified by (3) and (4). Here, the direct object of the English sentence corresponds to the grammatical subject of the German sentence, while the subject of the English sentence is realised as a prepositional phrase with *durch* in the German version.

(3) The conclusions of the Theato report safeguard them perfectly.[4]

(4) Durch die Schlußfolgerungen des
By the conclusions of_the
Berichts Theato werden sie
report Theato are they
uneingeschränkt bewahrt.
unlimitedly safeguarded

2.3 Negation

Sometimes, a positive predicate expression is translated by negating its antonym. This is the case in (5) and (6): both sentences contain a negative statement, but while the negation is incorporated into the English adjective by means of the negative prefix *in-*, it is achieved syntactically in the German sentence.

(5) the Directive is inapplicable in Denmark[5]

(6) die Richtlinie ist in Dänemark nicht
the Directive is in Denmark not
anwendbar
applicable

2.4 Information Structure

Sentences and their translations can be organised differently with regard to their information structure. Sentences (7) and (8) are a good example for this type of non-literal translation.

(7) Our motion will give you a great deal of food for thought, Commissioner[6]

(8) Eine Reihe von Anregungen werden
A row of suggestions will
wir Ihnen, Herr Kommissar, mit
we you, Mr. Commissioner, with
unserer Entschließung mitgeben
our resolution give

The German sentence is rather inconspicuous, with the grammatical subject being a prototypical agent (*wir*, 'we'). In the English version, however, it is the means that is realised in subject position and thus perspectivised. The corresponding constituent in German (*mit unserer Entschließung*, 'with our motion') is but an adverbial. In English, the actual agent is not realised as such and can only be identified by a process of inference based on the presence of the possessive pronoun *our*. Thus, while being more or less equivalent in meaning, this sentence pair differs significantly in its overall organisation.

ily discernible from the corpus data which language is the source language. Consequently, our use of the terms 'source', 'target', 'L1', and 'L2' does not admit of any conclusions as to whether one of the languages is the source language, and if so, which one.

[3] Europarl:de-en/ep-00-01-19.al, 489.

[4] Europarl:de-en/ep-00-01-18.al, 749.

[5] Europarl:de-en/ep-00-01-18.al, 2522.

[6] Europarl:de-en/ep-00-01-18.al, 53.

3 Alignment Model

The alignment model we propose is based on the assumption that a representation of translational equivalence can best be approximated by aligning the elements of monolingual predicate-argument structures. Section 3.1 describes this layer of the model in detail and shows how some of the differences in translations described in Section 2 can be accomodated on such a level. We assume that the annotation model described here is an extension to linguistic data which are already annotated with phrase-structure trees, i. e. treebanks. Section 3.2 shows how the binding of predicates and arguments to syntactic nodes is modelled. Section 3.3 describes the details of the alignment layer and the tags used to mark particular kinds of alignments, thus accounting for some more of the differences shown in Section 2.

3.1 Predicates and Arguments

The predicate-argument structures used in our model consist solely of predicates and their arguments. Although there is usually more than one predicate in a sentence, no attempt is made to nest structures or to join the predications logically in any way. The idea is to make the predicate-argument structure as rich as is necessary to be able to align a sentence pair while keeping it as simple as possible so as not to make it too difficult to annotate. In the same vein, quantification, negation, and other operators are not annotated. In short, the predicate-argument structures are not supposed to capture the semantics of a sentence exhaustively in an interlingua-like fashion.

To have clear-cut criteria for annotators to determine what a predicate is, we rely on the heuristic assumption that predicates are more likely to be expressed by tokens belonging to some word classes than by tokens belonging to others. Potential predicate expressions in this model are verbs, deverbal adjectives and nouns[7] or other adjectives and nouns which show a syntactic subcategorisation pattern. The predicates are represented by the capitalised citation form of the lexical item (e. g. HARMONISE). They are assigned a class based on their syntactic form (*v*, *n*, *a* for 'verbal', 'nominal', and 'adjectival', respectively), and derivationally related predicates form a predicate group.

Arguments are given short intuitive role names (e. g. ENT_HARMONISED, i. e. the entity being harmonised) in order to facilitate the annotation process. These role names have to be used consistently only within a predicate group. If, for example, an argument of the predicate HARMONISE has been assigned the role ENT_HARMONISED and the annotator encounters a comparable role as argument to the predicate HARMONISATION, the same role name for this argument has to be used.[8]

The usefulness of such a structure can be shown by analysing the sentence pair (1) and (2) in Section 2.1. While the syntactic constructions differ considerably, the predicate-argument structure shows the correspondence quite clearly (see the annotated sentences in Figure 1[9]): in the English sentence, we find the predicate HARMONISE with its argument ENT_HARMONISED, which corresponds to the predicate HARMONISIERUNG and its argument HARMONISIERTES in the German sentence. The information that a predicate of the class *v* is aligned with a predicate of the class *n* can be used to query the corpus for this type of non-literal translations.

The active vs. passive translation in sentences (3) and (4) is another phenomenon which is accomodated by a predicate-argument structure (Figure 2): the subject NP$_{502}$ in the English sentence corresponds to the passivised subject NP$_{502}$ (embedded in PP$_{503}$) in the German sentence on the basis of having the same argument role (SAFEGUARDER vs. BEWAHRER) in a comparable predication.

It is sometimes assumed that predicate-argument structure can be derived or recovered from constituent structure or functional tags such as subject and object.[10] It is true that these annotation layers provide important heuristic clues for the identification of predi-

[7] For all non-verbal predicate expressions for which a derivationally related verbal expression exists it is assumed that they are deverbal derivations, etymological counter-evidence notwithstanding.

[8] Keeping the argument names consistent for all predicates within a group while differentiating the predicates on the basis of syntactic form are complementary principles, both of which are supposed to facilitate querying the corpus. The consistency of argument names within a group, for example, enables the researcher to analyse paradigmatically all realisations of an argument irrespective of the syntactic form of the predicate. At the same time, the differentiation of predicates makes possible a syntagmatic analysis of the differences of argument structures depending on the syntactic form of the predicate.

[9] All figures are at the end of the paper.

[10] See e. g. (Marcus et al., 1994).

cates and arguments and may eventually speed up the annotation process in a semi-automatic way. But, as the examples above have shown, predicate-argument structure goes beyond the assignment of phrasal categories and grammatical functions, because the grammatical category of predicate expressions and consequently the grammatical functions of their arguments can vary considerably. Also, the predicate-argument structure licenses the alignment relation by showing explicitly what it is based on.

3.2 Binding Layer

As mentioned above, we assume that the annotation model described here is used on top of syntactically annotated data. Consequently, all elements of the predicate-argument structure must be bound to elements of the phrasal structure (terminal or non-terminal nodes). These bindings are stored in a dedicated binding layer between the constituent layer and the predicate-argument layer.

A problem arises when there is no direct correspondence between argument roles and constituents. For instance, this is the case whenever a noun is postmodified by a participle clause: in Figure 3, the argument role ENT_RAISED of the predicate RAISE is realised by NP$_{525}$, but the participle clause (IPA$_{517}$) containing the predicate (*raised*$_{6}$) needs to be excluded, because not excluding it would lead to recursion. Consequently, there is no simple way to link the argument role to its realisation in the tree.

In these cases, the argument role is linked to the appropriate phrase (here: NP$_{525}$) and the constituent that contains the predicate (IPA$_{517}$) is pruned out, which results in a discontinuous argument realisation. Thus, in general, the binding layer allows for complex bindings, with more than one node of the constituent structure to be included in and sub-nodes to be explicitly excluded from a binding to a predicate or argument.[11]

When an expected argument is absent on the phrasal level due to specific syntactic constructions, the binding of the predicate is tagged accordingly, thus accounting for the missing argument. For example, in passive constructions like in Table 1, the predicate binding is tagged as `pv`. Other common examples are imperative constructions. Although information of this kind may possibly be derived from the constituent structure, it is explicitly recorded in the binding layer as it has a direct impact on the predicate-argument structure and thus might prove useful for the automatic extraction of valency patterns.

Sentence	wenn	korrekt	gedolmetscht	wurde
Gloss	if	correctly	interpreted	was
			↑	
Binding			`pv`	
			│	
Pred/Arg			DOLMETSCHEN	

Table 1: Example of a tagged predicate binding (Europarl:de-en/ep-00-01-18.al, 2532)

Note that the passive tag can also be exploited in order to query for sentence pairs like (3) and (4) (in Section 2.2), where an active sentence is translated with a passive: it is straightforward to find those instances of aligned predicates where only one binding carries the passive tag.

3.3 Alignment Layer

On the alignment layer, the elements of a pair of predicate-argument structures are aligned with each other. Arguments are aligned on the basis of corresponding roles within the predications. Comparable to the tags used in the binding layer that account for specific constructions (see Section 3.2), the alignments may also be tagged with further information. These tags are used to classify types of non-literalness like those discussed in Sections 2.3 and 2.4.[12]

Sentences (5) and (6) are an example for a tagged alignment. As Section 2.3 has shown, negation may be incorporated in a predicate in L1, but not in L2. Since our predicate-argument structure does not include syntactic negation, this results in the alignment of a predicate in L1 with its logical opposite in L2. To account for this fact, predicate alignments of this kind are tagged as absolute opposites (`abs-opp`).

Similarly, alignment tagging is applied when predications are in some way incompatible, as is the case with sentences (7) and (8) in Section 2.4. As can be seen in the aligned annotation (Figure 4), the different information structure of these sentences has caused the two corresponding argument roles of GIVER and MITGEBER to be realised by two incompatible expressions representing different referents (NP$_{500}$

[11] See the database documentation (Feddes, 2004) for a more detailed description of this mechanism.

[12] The deviant translations described in Sections 2.1 and 2.2 are already represented via predicate class (see Section 3.1) and on the binding layer (see Section 3.2), respectively.

vs. wir_5). In this case, the alignment between the incompatible arguments is tagged `incomp`.

If there is no corresponding predicate-argument structure in the other language (as e. g. the adjectival predicate in sentence (2)) or if an argument within a structure does not have a counterpart in the other language, there will be no alignment.

Table 2 gives an overview of the annotation layers as described in this section.

Layer	*Function*
Phrasal	constituent structure of language A
Binding	binding ↓ predicates/arguments to ↑ nodes
PA	predicate-argument structures
Alignment	aligning ↕ predicates and arguments
PA	predicate-argument structures
Binding	binding ↑ predicates/arguments to ↓ nodes
Phrasal	constituent structure of language B

Table 2: The layers of the predicate-argument annotation

All elements of the alignment structure are supposed to mark explicitly the way they contribute to or distort the resulting translational equivalence of a sentence pair.[13] First and foremost, if two elements are aligned to each other, this alignment is licensed by their having comparable roles in the predicate-argument structures. This is the default case. If, however, a particular alignment relation, either of predicates or of arguments, is deviant in some way, this deviance is explicitly marked and classified on the alignment layer.

4 Application and Outlook

The alignment model we have described is currently being used in a project to build a treebank of aligned parallel texts in English and German with the following linguistic levels: POS tags, constituent structure and functional relations, plus the predicate-argument structure and the alignment layer to "fuse" the two – hence our working title for the treebank, FuSe, which additionally stands for *fu*nctional *se*mantic annotation (Cyrus et al., 2003; Cyrus et al., 2004).

Our data source, the Europarl corpus (Koehn, 2002), contains sentence-aligned proceedings of the European parliament in eleven languages and thus offers ample opportunity for extending the treebank at a later stage.[14] For syntactic and functional annotation we basically adapt the TIGER annotation scheme (Albert and others, 2003), making adjustments where we deem appropriate and changes which become necessary when adapting to English an annotation scheme which was originally developed for German.

We use Annotate for the semi-automatic assignment of POS tags, hierarchical structure, phrasal and functional tags (Brants, 1999; Plaehn, 1998a). Annotate stores all annotations in a relational database.[15] To stay consistent with this approach we have developed an extension to the Annotate database structure to model the predicate-argument layer and the binding layer.

Due to the monolingual nature of the Annotate database structure, the alignment layer (Section 3.3) cannot be incorporated into it. Hence, additional types of databases are needed. For each language pair (currently English and German), an alignment database is defined which represents the alignment layer, thus fusing two extended Annotate databases. Additionally, an administrative database is needed to define sets of two Annotate databases and one alignment database. The final parallel treebank will be represented by the union of these sets (Feddes, 2004).

While annotators use Annotate to enter phrasal and functional structures comfortably, the predicate-argument structures and alignments are currently entered into a structured text file which is then imported into the database. A graphical annotation tool for these layers is under development. It will make binding the predicate-argument structure to the constituent structure easier for the annotators and suggest argument roles based on previous decisions.

Possiblities of semi-automatic methods to speed up the annotation and thus reduce the costs of building the treebank are currently being investigated.[16] Still, quite a bit of manual

[13] Cf. the "translation network" described in (Santos, 2000) for a much more complex approach to describing translation in a formal way; this model, however, goes well beyond what we think is feasible when annotating large amounts of data.

[14] There are a few drawbacks to Europarl, such as its limited register and the fact that it is not easily discernible which language is the source language. However, we believe that at this stage the easy accessibility, the amount of preprocessing and particularly the lack of copyright restrictions make up for these disadvantages.

[15] For details about the Annotate database structure see (Plaehn, 1998b).

[16] One track we follow is to investigate if it is feasible to

work will remain. We believe, however, that the effort that goes into such a gold-standard parallel treebank is very much worthwhile since the treebank will eventually prove useful for a number of fields and can be exploited for numerous applications. To name but a few, translation studies and contrastive analyses will profit particularly from the explicit annotation of translational differences. NLP applications such as Machine Translation could, e. g., exploit the constituent structures of two languages which are mapped via the predicate-argument-structure. Also, from the disambiguated predicates and their argument structures, a multilingual valency dictionary could be derived.

References

Stefanie Albert et al. 2003. TIGER Annotationsschema. Technical report, Universität des Saarlandes, Universität Stuttgart, Universität Potsdam. Unpublished Draft – 24 July 2003.

Thorsten Brants. 1999. *Tagging and Parsing with Cascaded Markov Models: Automation of Corpus Annotation*, volume 6 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology.* Saarland University, Saarbrücken.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2003. FuSe – a multi-layered parallel treebank. Poster presented at the Second Workshop on Treebanks and Linguistic Theories, 14–15 November 2003, Växjö, Sweden (TLT 2003). `http://fuse.uni-muenster.de/Publications/0311_tltPoster.pdf`.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2004. Annotating predicate-argument structure for a parallel treebank. In Charles J. Fillmore, Manfred Pinkal, Collin F. Baker, and Katrin Erk, editors, *Proc. LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon, May 30, 2004*, pages 39–46. `http://fuse.uni-muenster.de/Publications/0405_lrec.pdf`.

Hendrik Feddes. 2004. FuSe database structure. Technical report, Arbeitsbereich Linguistik, University of Münster. `http://fuse.uni-muenster.de/Publications/dbStruktur.pdf`.

Silvia Hansen-Schirra and Stella Neumann. 2003. The challenge of working with multilingual corpora. In Stella Neumann and Silvia Hansen-Schirra, editors, *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives. Corpus Linguistics 2003, Lancaster*, pages 1–6.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished draft, `http://www.isi.edu/~koehn/publications/europarl/`.

Mitch Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proc. ARPA Human Language Technology Workshop.*

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-07, IRCS, University of Pennsylvania. `http://citeseer.ist.psu.edu/melamed98manual.html`.

Oliver Plaehn. 1998a. ANNOTATE Bedienungsanleitung. Technical report, Universität des Saarlandes, FR 8.7, Saarbrücken. `http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate-manual.ps.gz`.

Oliver Plaehn. 1998b. ANNOTATE Datenbank-Dokumentation. Technical report, Universität des Saarlandes, FR 8.7, Saarbrücken. `http://www.coli.uni-sb.de/sfb378/negra-corpus/datenbank.ps.gz`.

Raphael Salkie. 2002. How can linguists profit from parallel corpora? In Lars Borin, editor, *Parallel Corpora, Parallel Worlds*, pages 93–109. Rodopi, Amsterdam.

Diana Santos. 2000. The translation network: A model for a fine-grained description of translations. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, volume 13 of *Text, Speech and Language Technology*, chapter 8. Kluwer, Dordrecht.

have the annotators mark predicate-argument structures on raw texts and have the phrasal and functional layers added in a later stage, possibly supported by methods which derive these layers partially from the predicate-argument structures. This is, however, still very tentative.

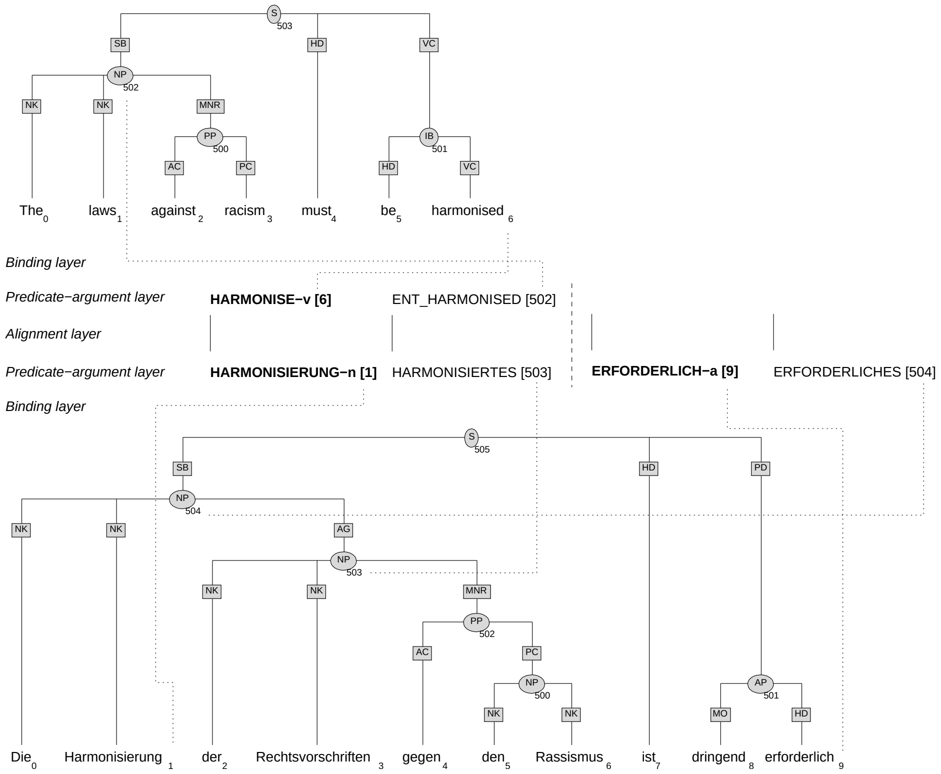

Figure 1: Alignment of a verb/direct-object construction with a noun/modifier construction

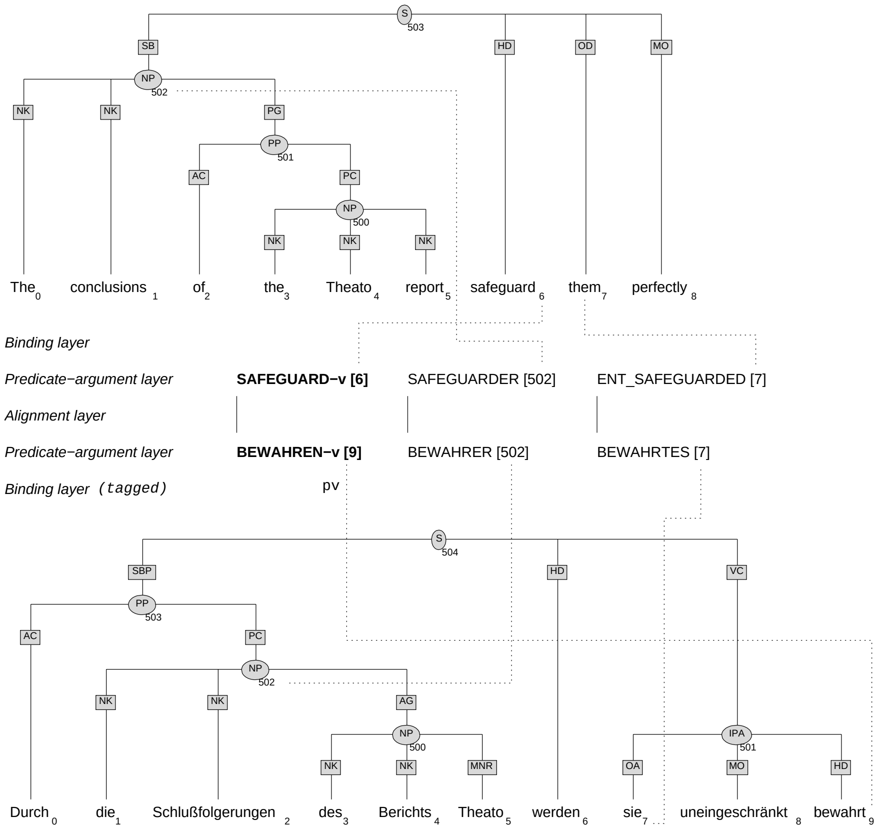

Figure 2: Active vs. passive voice in translations: an example of a tagged binding (pv)

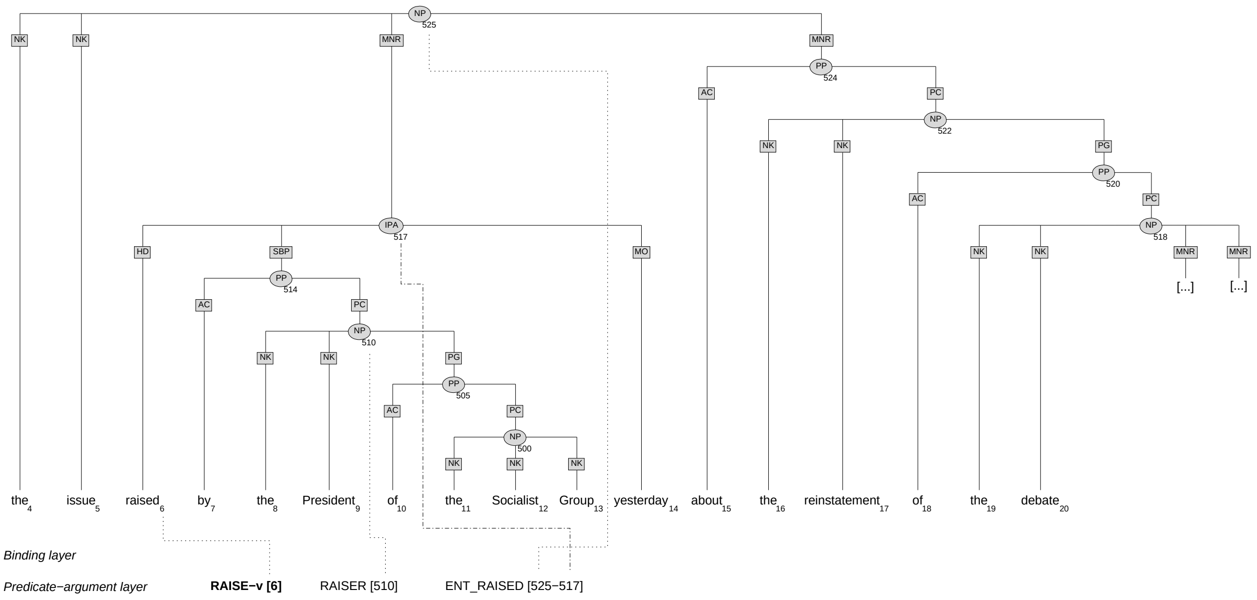

Figure 3: Complex binding of an argument: an example of a pruned constituent (dash-dotted line)

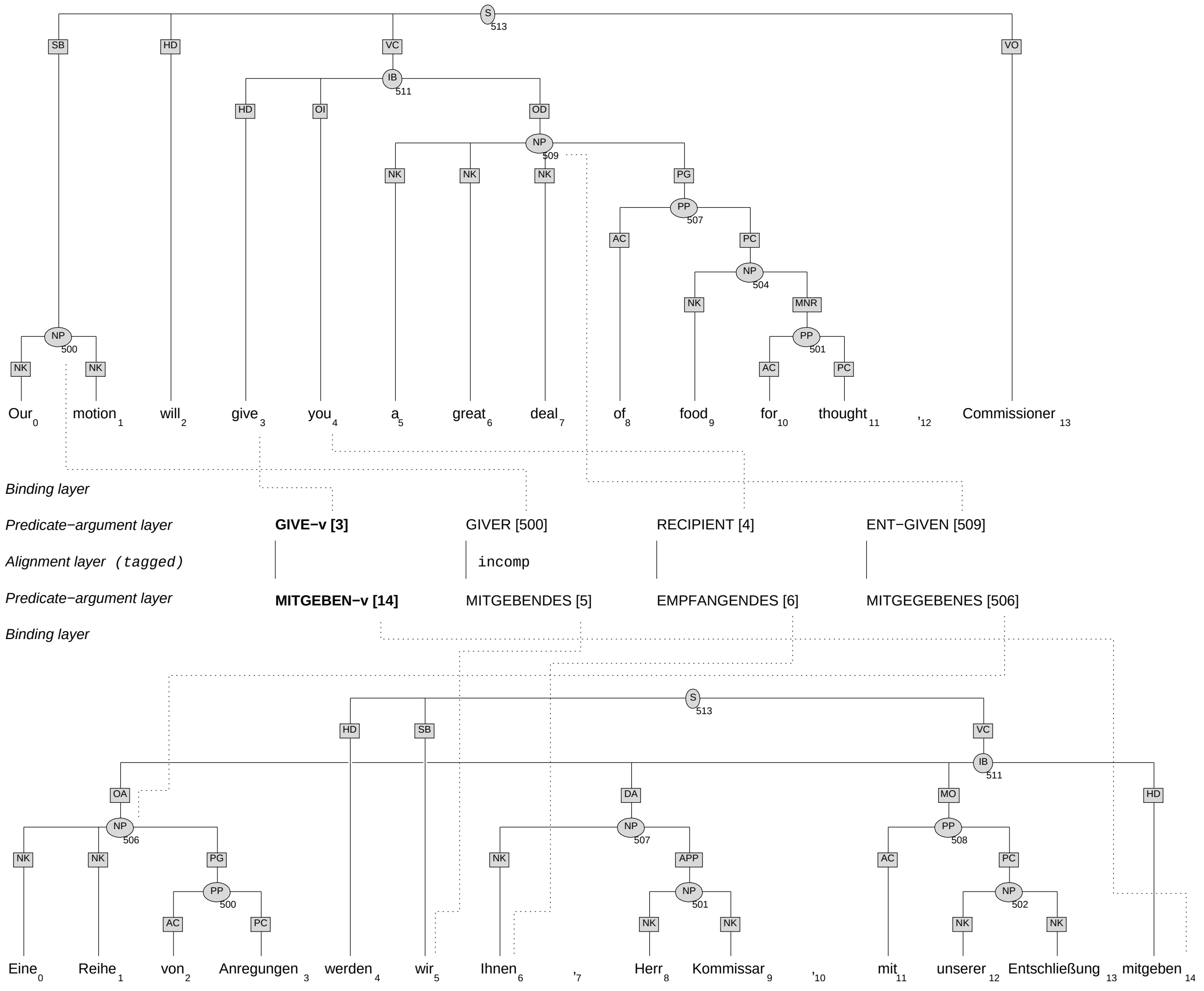

Figure 4: Different information structure: an example of a tagged alignment (`incomp`)