

Phrase Alignment in Parallel Treebanks

Yvonne Samuelsson and Martin Volk

Stockholm University
Department of Linguistics

E-mail: {yvonne.samuelsson, volk}@ling.su.se

1 Introduction

The combined research on treebanks and parallel corpora has recently led to parallel treebanks. A parallel treebank consists of syntactically annotated sentences in two or more languages, taken from translated (i.e. parallel) documents. In addition, the syntax trees of two corresponding sentences are aligned on a sub-sentential level (word, phrase and clause level). Parallel treebanks can be used as training or evaluation corpora for word and phrase alignment, as input for example-based machine translation (EBMT), as training corpora for transfer rules, or for translation studies.

We are developing a German-English-Swedish parallel treebank, consisting of over 1000 sentences in each language. This paper is a report on experiences regarding the alignment. We will look at the tools, the alignment guidelines and the inter-annotator agreement.

2 Building the Treebanks

Our parallel treebank contains the first two chapters of Jostein Gaarder's novel "Sofie's World" (the original is the Norwegian, [4]). This part contains around 530 sentences in each language (there is some variation between the different language versions), with an average of about 14 tokens per sentence. The second part of the parallel treebank contains economy texts, taken from a quarterly report by a multinational company, a bank's annual report and a text about a banana certification program. This part contains around 490 sentences, with an average of about 22 tokens per sentence. Besides having more tokens per sentence, the economy texts are also more complex and differ more in number of sentences and average number of tokens between the languages.

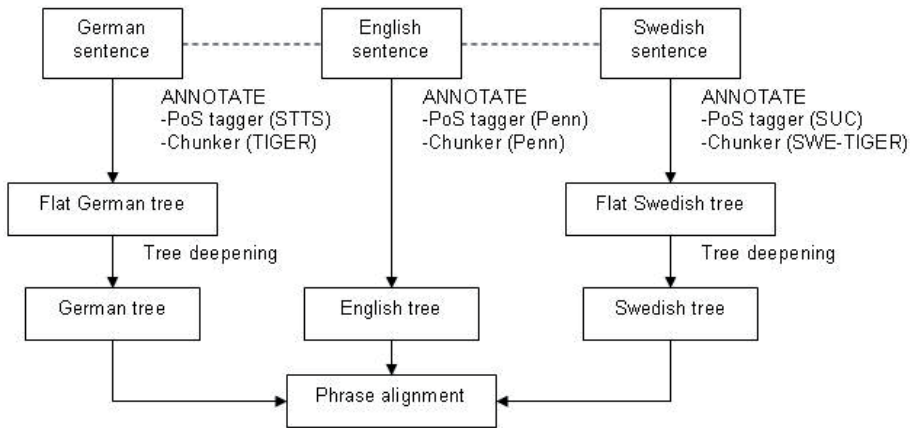


Figure 1: The process of creating the parallel treebank, step-by-step.

Figure 1 shows our work flow for the creation of the parallel treebank. First, we annotated the monolingual treebanks with the ANNOTATE treebank editor¹. It includes Thorsten Brants' statistical Part-of-Speech Tagger and Chunker. The first annotation step was to automatically tag the sentences with Part-of-Speech tags. For the English treebank we used the Penn Treebank Part-of-Speech tag set while the German is annotated with the STTS (Stuttgart-Tübingen Tag Set [8]), and the Swedish with an adapted version of the SUC tag set (Stockholm-Umeå Corpus). We then semi-automatically parsed the English sentences according to the Penn Treebank grammar, [1], while the German follows the TIGER annotation schema, [7, 2]. For the Swedish treebank we used an adapted version of the German TIGER guidelines. This adaptation is tailored to account for specific Swedish constructions and problems, such as using the function labels DO (direct object) and IO (indirect object) instead of the German OA (accusative object) and DA (dative), or the fact that Swedish prepositional phrases can consist of a preposition plus a sentence or verb phrase.

As we will see later on, the use of different annotation schemata for different languages is sometimes problematic for the alignment. However, we want the monolingual treebanks to be standalone, in addition to being used together in the parallel treebank, and therefore compatible with existing treebanks.

The Penn and TIGER formats differ in many respects. In table 1 we see that the TIGER format is shallower when it comes to e.g. noun phrases. While the

¹www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

Penn treebank	TIGER treebank
<pre> graph TD NP1[NP] --- NP2[NP] NP1 --- PP[PP] NP2 --- Det[Det] NP2 --- N[N] </pre>	<pre> graph TD NP1[NP] --- Det1[Det] NP1 --- N1[N] NP1 --- PP[PP] </pre>
<pre> graph TD S[S] --- NPsubj[NP_{Subj}] S --- VP1[VP] VP1 --- Vfin[V_{Fin}] VP1 --- VP2[VP] VP2 --- Vinf[V_{Inf}] VP2 --- NPobj[NP_{Obj}] </pre>	<pre> graph TD S[S] --- NPsubj[NP_{Subj}] S --- Vfin[V_{Fin}] S --- VP[VP] VP --- NPobj[NP_{Obj}] VP --- Vinf[V_{Inf}] </pre>

Table 1: Some structural differences between the English Penn treebank and the German TIGER treebank.

TIGER NP dominates determiner, noun and modifying PP directly, the Penn NP first splits into an NP and a PP, to have the determiner and noun in an NP of its own. We also see that the Penn sentence (S) directly dominates the subject and a VP, which in turn contains the finite verb. The TIGER sentence, on the other hand, directly dominates the subject and the finite verb while VP’s are reserved for infinite verbs. Additionally, the Penn treebank annotation contains traces (empty tokens) while the TIGER annotation allows for crossing branches. The TIGER annotation requires function labels on every edge while they are sparse in the Penn annotation, only used for specific functions like subject, predicate and different types of adjuncts (temporal, local, manner, etc.).

The TIGER annotation guidelines thus give a flat phrase structure tree without unary nodes, “unnecessary” NPs (noun phrases) within PPs (prepositional phrases) and finite VPs (verb phrases). Using a flat tree structure means fewer annotation decisions, and a better overview of the trees for the human annotator. However, it also means that the trees are not complete from a linguistic point of view. Moreover, flat syntax trees are problematic for the phrase alignment since we want to be able to draw the alignment on as many levels as possible. Therefore, we deepen the German and Swedish trees automatically with a program, which inserts unambiguous nodes, like an AP (adjective phrase) for the adjective in an NP, and an NP in flat PP’s (prepositional phrases). This procedure is described in detail in [6]. We do not deepen the English structures. However, we realize that we could speed up the an-

notation process if we would annotate flatter English trees, which are then automatically deepened. Furthermore the Penn guidelines advocate flat structures within the core NP (Det Adj* Noun+), which would profit from internal groupings, such as AP's or noun groups.

3 Alignment

After creating the monolingual treebanks, we converted the trees into TIGER-XML, a powerful database-oriented representation for graph structures². In a TIGER-XML graph each leaf (= token) and each node (= linguistic constituent) has a unique identifier (prefixed with the sentence number). We use these unique node identifiers for the phrase alignment across trees in corresponding translation units. We also use an XML representation for storing this alignment.

Phrase alignment can be regarded as an additional layer of information on top of the syntax structure. It shows which part of a sentence in one language is equivalent to which part of a corresponding sentence in another language. We draw alignment lines manually between sentences, phrases and words over parallel trees with the help of the Stockholm TreeAligner, a graphical user interface to insert (or correct) alignments between pairs of syntax trees. This tool has been described in [9]. Figure 2 shows an example of two aligned trees. Alignment is drawn as lines between words and nodes. We see that e.g. the English word *answer* corresponds to the Swedish word *svar* and that the English NP dominating *she* corresponds to the Swedish NP *hon*.

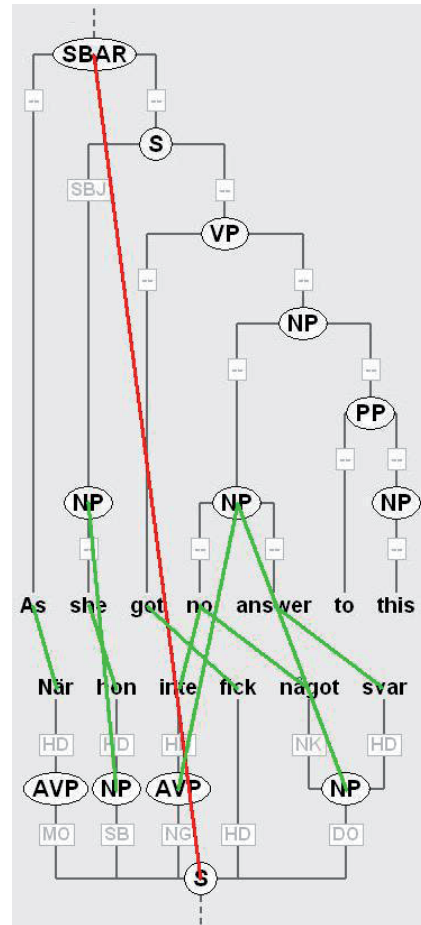


Figure 2: Two aligned trees (partial), English (*As she got no answer to this*) and Swedish (*När hon inte fick något svar, as she not got any answer*).

²See <http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html>.

Currently the TreeAligner provides for manual alignment on words and nodes. We have experimented with automatic word alignment and its projection to node alignments. The idea is to use statistical word alignment methods to predict word correspondences and to project these correspondences via the phrase heads (which are marked in the trees) to the respective nodes. This approach looks very promising, but is not yet implemented in the TreeAligner.

3.1 Alignment Guidelines

We want to align as many phrases as possible. The goal is to show translation equivalence, focusing on meaning rather than sentence structure. Phrases shall only be aligned if the tokens that they span represent the same meaning and if they could serve as translation units outside the current sentence context. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit.

In the following we will mostly talk about node (as opposed to word) alignment, node and word alignment being the two types of phrase alignment. Usually alignment is word-to-word or node-to-node. (In the following examples [brackets]NP denote nodes, EN is English, DE German and SV Swedish.)

We have two types of alignment, displayed by different colours in our alignment tool. Nodes/words representing exactly the same meaning are aligned as exact translation correspondences, like in example 1. If they represent approximately the same meaning, they are aligned as fuzzy translation correspondences, like in example 2, because of the pronoun *her*.

- (1) **DE:** [*den mänskliga hjärnan*]NP
EN: [*the human brain*]NP

- (2) **DE:** [*auf dem Heimweg von der Schule*]PP
(*on the home-way from the school*)
EN: [*on her way home from school*]PP

Our alignment guidelines allow phrase alignments within m:n sentence alignments. Even though m:n phrase alignments are technically possible, we have only used 1:n phrase alignments (not specifying the direction), for simplicity and clarity reasons. The 1:n alignment option is not used if a node from one tree is realized twice in the corresponding tree. Example 3 shows 1:n alignment on the word level. An example of 1:n alignment on the node level can be seen in figure 2 where the English node containing *no answer* is aligned to two Swedish nodes containing *inte* (not) and *något svar* (any answer).

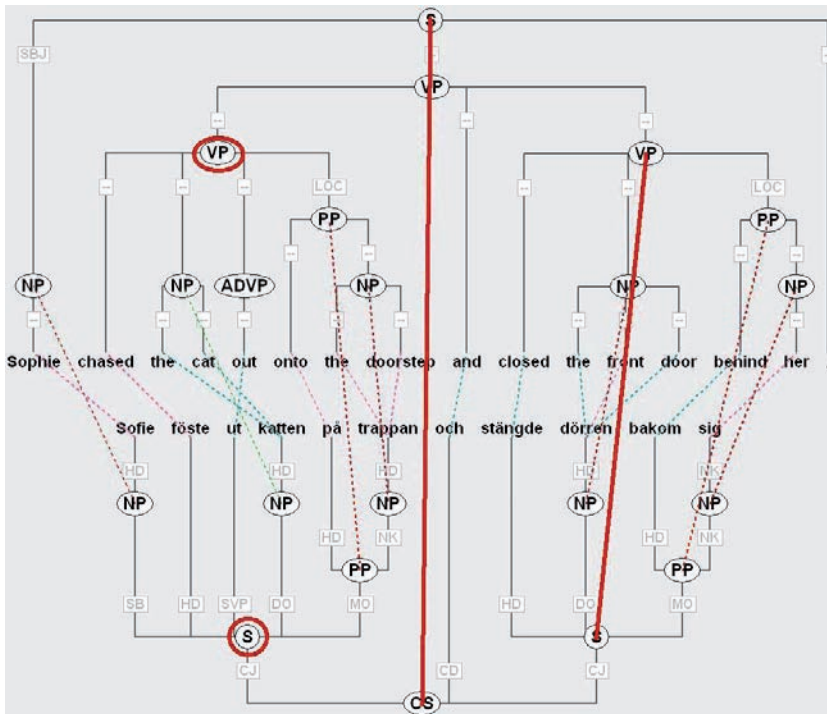


Figure 3: Structural differences between the annotation schemata leave the marked nodes without alignment.

- (3) **SV:** *fruktträden*
EN: *the fruit trees*

Pronouns should not be aligned to full noun phrases. Nodes that contain extra information in one language cannot be aligned. Figure 3 shows two sentences, the Swedish containing two coordinated sentences (one without subject) and the English containing two coordinated verb phrases. The Swedish coordinated sentence is aligned to the English S, and the second Swedish S is aligned to the second English inner VP (marked with a thick line). The first Swedish S however cannot be aligned to the first English VP (both nodes marked with a thick ring) since the Swedish node contains the subject. This problem is frequent in our treebank, due to differences in the annotation schemata between the languages.

Another example of the problem with extra information in one language is example 4 where the Swedish NP contains a relative clause, meaning that it cannot

	A1	A2	A3
Node alignments	506 (42%)	610 (39%)	603 (39%)
Word alignments	690 (58%)	943 (61%)	952 (61%)
Total	1196	1553	1555

Table 2: Number and type (node or word) of alignment links (percentage of total number of alignments by each annotator).

be aligned to the English NP.

- (4) **SV:** *[det [raka]AP håret [som...]S]NP*
(the straight hair which...)
EN: *[her straight hair]NP*

There are of course many problematic cases, some due to the translator’s freedom. This can be seen on all levels (word, phrase, clause and sentence). In these cases it is important to remember the main goal, that the aligned phrases should be equivalent in meaning outside the sentence context.

3.2 Inter-Annotator Agreement

How reliable is the manual alignment? How comprehensive are our alignment guidelines and are there questions not addressed? In an attempt to answer these questions, and to perfect the annotation guidelines, we carried out a small evaluation of inter-annotator agreement. In addition to the student annotator, both authors, independently of each other, also aligned one hundred sentence pairs of the English-Swedish treebank (the hundred first sentence “pairs” of the Sofie treebanks, English sentences 1-100, Swedish sentences 1-103). Thus we had three versions of the alignments to compare. (A1 is annotator 1, etc.) A Perl program was created which compares the alignment files.

Table 2 shows the annotation for each annotator. Annotators 2 and 3 have approximately the same amount of alignment links while annotator 1 has been more restrictive, both on the word and node levels.

In table 3 we see the agreement when comparing the annotations. Partial agreement means that the annotators agree that a phrase pair should be aligned, but they have different opinions about whether it should be exact or fuzzy alignment. The percentages are computed as the intersection of alignments divided by all the alignments of the annotator. For example, if we compare annotators 1 and 2, they fully

	Number	A1	A2	A3	
A1-A2-A3	882	74%	57%	57%	Full agreement
	163	14%	10%	10%	Partial agreement
	1045	87%	67%	67%	Total (full or partial)
A1-A2	971	81%	63%	-	Full agreement
	125	10%	8%	-	Partial agreement
	1096	92%	71%	-	Total (full or partial)
A1-A3	986	82%	-	63%	Full agreement
	121	10%	-	8%	Partial agreement
	1107	93%	-	71%	Total (full or partial)
A2-A3	1168	-	75%	75%	Full agreement
	186	-	12%	12%	Partial agreement
	1354	-	87%	87%	Total (full or partial)

Table 3: Agreement comparing all three and two annotators at a time (percentage of total number of alignments by each annotator).

agree on 971 alignments. These 971 alignments are 81% of 1196 (see table 2), the total number of alignments for annotator 1, and 63% of 1553, the total number of alignments for annotator 2.

The percentage of agreement for the alignment of annotator 1 is higher than the others, presumably since annotator 1 has fewer alignment links. Because the scores are so high, one might think that the task was too easy. However, since the scores get lower for annotators 2 and 3, it is more likely that annotator 1 has handled the most obvious alignments, leaving many of the more difficult problems unaligned.

One might regard the agreement as low, when comparing all three annotators. Let us keep in mind that the alignment task is more comparable to semantic annotation than to PoS tagging or parsing. The SALSA project, dealing with semantic frame annotation, reported inter-annotator agreement of 85-86% when comparing two annotations [3]. Thus our alignment guidelines still need to be refined. However, it is interesting that the figures for full agreement are so high, compared to the figures for partial agreement. This means that when there is agreement on which words/nodes to align, then there is often no problem deciding whether it should be exact or fuzzy alignment.

Let us look at the distribution of nodes and words in the alignments agreed upon by all three annotators, table 4. Interestingly enough, the word alignment gets better scores when it comes to full agreement while node alignment gets better scores

Nodes	Words	
332 (38%)	550 (62%)	Full agreement
92 (56%)	71 (44%)	Partial agreement
424 (41%)	621 (59%)	Total (full and partial)

Table 4: Node and word alignment agreed upon by all annotators.

in partial agreement. This means that the annotators more often agree on word alignment, and when they agree on node alignment they more often disagree about whether the alignment should be exact or fuzzy. This is not so surprising, considering that a node often contains more information than just one word. Therefore, it is more difficult to set the boundary for when something is an exact translation rather than an approximate translation.

It is also interesting to note that annotator 1 used exact alignment for 79% of the links, annotator 2 used exact alignment for 73% of the links and annotator 3 for 69%. In the partial agreement annotators 1 and 2 were alone in their decision of exact alignment more often than in their decision of fuzzy alignment (31% against 20% for annotator 1 and 38% against 27% for annotator 2), while annotator 3 more often was alone in choosing fuzzy alignment (30% against 17%).

In an alignment project like this, with the large total number of nodes and words, it is easy to miss alignments that should be drawn. This could be solved by forcing the annotator to mark phrases without correspondence as unaligned (as was proposed in the Blinker project, [5]). However, the vast amount of extra time it would take to mark everything without correspondence, both on word and on node level, made this approach unmanageable.

In table 5 we see the alignment only found in one of the three annotations. Annotator 1 had only a few annotations that none of the others agreed upon. This again points to the conclusion that annotator 1 has chosen to do safer alignments, leaving more difficult cases unaligned. Looking at the actual annotation, the alignments only found in the file by annotator 1 are mostly errors, like aligning the Swedish word *Sofie* to the English noun phrase (NP) containing the word *Sophie* (it should be word-to-word alignment) or aligning the Swedish NP containing *Kaptenssvängen* to the English NP containing *Captain's* when it should be the larger English NP *Captain's Bend*. Annotator 1 has more node than word alignment which the others do not agree upon. This is however explained by the fact that the few word-to-node alignments are counted as node alignments by our comparison program and several of the alignment links here are of the same type as the *Sophie*-example

A1	A2	A3	
38 (3%)	148 (10%)	139 (9%)	No agreement
28 (6%)	65 (11%)	44 (7%)	No node agreement
10 (1%)	83 (9%)	95 (10%)	No word agreement

Table 5: Alignment not agreed upon, present only with one annotator (percentage of total number of alignments by each annotator).

just mentioned.

Annotator 2 has several times tried to align a sequence of words in one language to a sequence in the other language by drawing lines between all the words, e.g. between *was like* and *påminde om* (reminded of). This type of alignment makes the matching process more complex, when using the data in an EBMT system. If we return to the 1:n alignment in example 3, finding *the* would give us *fruktträden*, we then would have to check the rest of the words aligned to *fruktträden* and only use the match if *the* is followed by *fruit trees*. With m:n alignment, using the example of *was like* above, we would need to check the context of both languages (*was* followed by *like* and *påminde* followed by *om*). Additionally, m:n alignment might mislead the annotator to align too many words. These are the main reasons to only allow 1:n alignment on word/node level.

Most alignments only present in the annotations of annotator 3 seem to come from the fact that this annotator has tried to align parts that are similar (rather than equivalent) in the translations. This means drawing word alignment between the adjectives in *very distant* and *ganska avlägsen* (rather distant), or between *asked* and *ställt* (put) in the phrases *asked the question* and *ställt frågan* (literally put the question). This means that the guidelines need to stress the main goal of the alignment even more, that the aligned parts should be equivalent outside the current sentence context.

3.3 Ensuring Alignment Quality

Based on the lessons learned in the inter-annotator agreement experiments, we improved our alignment guidelines. But how can we ensure that the guidelines are followed? We would like to determine whether the alignments are complete and consistent, in similarity to quality checks over treebanks. The completeness check will be difficult unless we require our annotators to use each word and each node of every tree. We could check if a certain sequence of tokens is aligned in some sentences and not in others, but we have not done that yet.

Instead we have started to work on consistency checking of the alignments. We check for all aligned single tokens and all aligned token sequences whether they are aligned in the same way (i.e. with the predicate 'exact' or 'fuzzy') to the same corresponding tokens. We also check whether the aligned token sequences differ in length (calculated as number of characters). Large length differences might point to erroneous alignments. Finally we examine those cases where different types of nodes are aligned across the languages (e.g. when an adjective phrase in one language is aligned with a prepositional phrase in the other).

These consistency checks are done manually over an extracted table of the aligned token sequences (with their node labels). This allows us to sort the token sequences according to different criteria and to abstract away from the dense forest of syntactic information and alignment lines in the TreeAligner.

In the future we plan to add checks that exploit dependencies between alignments. For example, if a node in language 1 is aligned to a node in language 2 and both nodes have exactly one daughter node, then these daughter nodes should also be aligned.

4 Conclusion

Much previous work on (alignment) annotation has used a gold standard to evaluate the annotation. In our case a gold standard would not have been of much help since comparing the three alignment files has mainly been important in creating the guidelines for the alignment. Looking at inter-annotator agreement is a clear way of finding the problematic cases that one annotator might not even be aware of as problematic.

Another problem with creating a gold standard is the fact that we are dealing with a kind of semantic annotation. Creating a gold standard for PoS tagging or parsing is easy in comparison, if there are detailed guidelines to follow. When dealing with meaning, a large part of the annotation is always open for discussion. It is not easy to decide about right and wrong.

The inter-annotator comparison shows that it is easy to miss alignment. One solution would be to force the annotator(s) to align everything, marking words/nodes without equivalence in the other language as unaligned. This would however be very time consuming. The agreement also shows that word alignment is usually easier, most likely because node alignment contains a higher degree of subjectivity, if two nodes are equivalent or not. Finally, the rate of full agreement is much higher than the rate of partial agreement (when the annotators do not agree on the type of alignment). This shows that when the annotators agree on which words/nodes to align, there is often no problem deciding whether it should be exact or fuzzy align-

ment.

Computing the inter-annotator agreement has given us many insights regarding problem areas in alignment annotation. This has led to an improvement of the alignment guidelines. We are, however, convinced that the guidelines need to be enhanced even further, as our work with the alignment proceeds.

References

- [1] A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. Bracketing guidelines for treebank II style, Penn treebank project. Technical report, University of Pennsylvania, 1995.
- [2] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.
- [3] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of the Conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genoa, 2006.
- [4] J. Gaarder. *Sofies verden: Roman om filosofiens historie*. Aschehoug, 1991.
- [5] I. Melamed. Manual annotation of translational equivalence: The Blinker project. Technical Report IRCS 98-07, Department of Computer and Information Science, University of Pennsylvania, 1998.
- [6] Y. Samuelsson and M. Volk. Automatic node insertion for treebank deepening. In *Proc. of 3rd Workshop on Treebanks and Linguistic Theories*, Tübingen, December 2004.
- [7] W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 88–95, Washington, DC, 1997.
- [8] C. Thielen, A. Schiller, S. Teufel, and C. Stöckert. Guidelines für das Tagging Deutscher Textkorpora mit STTS. Technical report, IMS and Sfs, 1999.
- [9] M. Volk, S. Gustafson-Capková, J. Lundborg, T. Marek, Y. Samuelsson, and F. Tidström. XML-based Phrase Alignment in Parallel Treebanks. In *Proc. of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, April 2006.