Human Judgements in Parallel Treebank Alignment

Martin Volk and Torsten Marek
University of Zurich
Institute of Computational Linguistics
8050 Zurich, Switzerland
volk@cl.uzh.ch

Yvonne Samuelsson
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
yvonne.samuelsson@ling.su.se

Abstract

We have built a parallel treebank that includes word and phrase alignment. The alignment information was manually checked using a graphical tool that allows the annotator to view a pair of trees from parallel sentences. We found the compilation of clear alignment guidelines to be a difficult task. However, experiments with a group of students have shown that we are on the right track with up to 89% overlap between the student annotation and our own. At the same time these experiments have helped us to pin-point the weaknesses in the guidelines, many of which concerned unclear rules related to differences in grammatical forms between the languages.

1 Introduction

Establishing translation correspondences is a difficult task. This task is traditionally called alignment and is usually performed on the paragraph level, sentence level and word level. Alignment answers the question: Which part of a text in language L1 corresponds in meaning to which part of a text in language L2 (under the assumption that the two texts represent the same meaning in different languages). This may mean that one text is the translation of the other or that both are translations derived from a third text.

There is considerable interest in automating the alignment process. Automatic sentence alignment of legacy translations helps to fill translation memories. Automatic word alignment is a crucial step in training statistical machine translation systems. Both sentence and word alignment have to deal with 1:many alignments, i.e. sometimes a sentence in one language is translated as two or three sentences in the other language.

In other respects sentence alignment and word alignment are fundamentally different. It is relatively safe to assume the same sentence order in both languages when computing sentence alignment. But such a monotonicity assumption is not possible for word alignment which needs to allow for word order differences and thus for crossing alignments. And while algorithms for sentence alignment usually focus on length comparisons (in terms of numbers of characters), word alignment algorithms use cross-language cooccurrence frequencies as a key feature.

Our work focuses on word alignment and on an intermediate alignment level which we call phrase alignment. Phrase alignment encompasses the alignment from simple noun phrases and prepositional phrases all the way to complex clauses. For example, on the word alignment level we want to establish the correspondence of the German "verb form plus separated prefix" *fing an* with the English verb form *began*. While in phrase alignment we mark the correspondence of the verb phrases *ihn in den Briefkasten gesteckt* and *dropped it in the mail box*.

We regard phrase alignment as alignment between linguistically motivated phrases, in contrast to some work in statistical machine translation where phrase alignment is defined as the alignment between arbitrary word sequences. Our phrase alignment is alignment between nodes in constituent structure trees. See figure 1 for an ex-

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 51–57
Manchester, August 2008

ample of a tree pair with word and phrase alignment.

We believe that such linguistically motivated phrase alignment provides useful phrase pairs for example-based machine translation, and provides interesting insights for translation science and cross-language comparisons. Phrase alignments are particularly useful for annotating correspondences of idiomatic or metaphoric language use.

2 The Parallel Treebank

We have built a trilingual parallel treebank in English, German and Swedish. The treebank consists of around 500 trees from the novel Sophie's World and 500 trees from economy texts (an annual report from a bank, a quarterly report from an international engineering company, and the banana certification program of the Rainforest Alliance). The sentences in Sophie's World are relatively short (14.8 tokens on average in the English version), while the sentences in the economy texts are much longer (24.3 tokens on average; 5 sentences in the English version have more than 100 tokens).

The treebanks in English and German consist of constituent structure trees that follow the guidelines of existing treebanks, the NEGRA/TIGER guidelines for German and the Penn treebank guidelines for English. There were no guidelines for Swedish constituent structure trees. We have therefore adapted the German treebank guidelines for Swedish. Both German trees and Swedish trees are annotated with flat structures but subsequently automatically deepened to result in richer and linguistically more plausible tree structures.

When the monolingual treebanks were finished, we started with the word and phrase alignment. For this purpose we have developed a special tool called the Stockholm TreeAligner (Lundborg et al., 2007) which displays two trees and allows the user to draw alignment lines by clicking on nodes and words. This tool is similar to word alignment tools like ILink (Ahrenberg et al., 2003) or Cairo (Smith and Jahr, 2000). As far as we know our tool is unique in that it allows the alignments of linguistically motivated phrases via node alignments in parallel constituent structure trees (cf. (Samuelsson and Volk, 2007)).

After having solved the technical issues, the challenge was to compile precise and comprehensive guidelines to ensure smooth and consistent alignment decisions. In (Samuelsson and Volk, 2006) we have reported on a first experiment to evaluate inter-annotator agreement from our alignment tasks.

In this paper we report on another recently conducted experiment in which we tried to identify the weaknesses in our alignment guidelines. We asked 12 students to alignment 20 tree pairs (English and German) taken from our parallel treebank. By comparing their alignments to our Gold Standard and to each other we gained valuable insights into the difficulty of the alignment task and the quality of our guidelines.

3 Related Research

Our research on word and phrase alignment is related to previous work on word alignment as e.g. in the Blinker project (Melamed, 1998) or in the UPLUG project (Ahrenberg et al., 2003). Alignment work on parallel treebanks is rare. Most notably there is the Prague Czech-English treebank (Kruijff-Korbayová et al., 2006) and the Linköping Swedish-English treebank (Ahrenberg, 2007). There has not been much work on the alignment of linguistically motivated phrases. Tinsley et al. (2007) and Groves et al. (2004) report on semi-automatic phrase alignment as part of their research on example-based machine translation.

Considering the fact that the alignment task is essentially a semantic annotation task, we may also compare our results to other tasks in semantic corpus annotation. For example, we may consider the methods for resolving annotation conflicts and the figures for inter-annotator agreement in frame-semantic annotation as found in the German SALSA project (cf. (Burchardt et al., 2006)).

4 Our Alignment Guidelines

We have compiled alignment guidelines for word and phrase alignment between annotated syntax trees. The guidelines consist of general principles, concrete rules and guiding principles.

The most important general principles are:

1. Align items that can be re-used as units in a machine translation system.
2. Align as many items (i.e. words and phrases) as possible.
3. Align as close as possible to the tokens.

The first principle is central to our work. It defines the general perspective for our alignment.

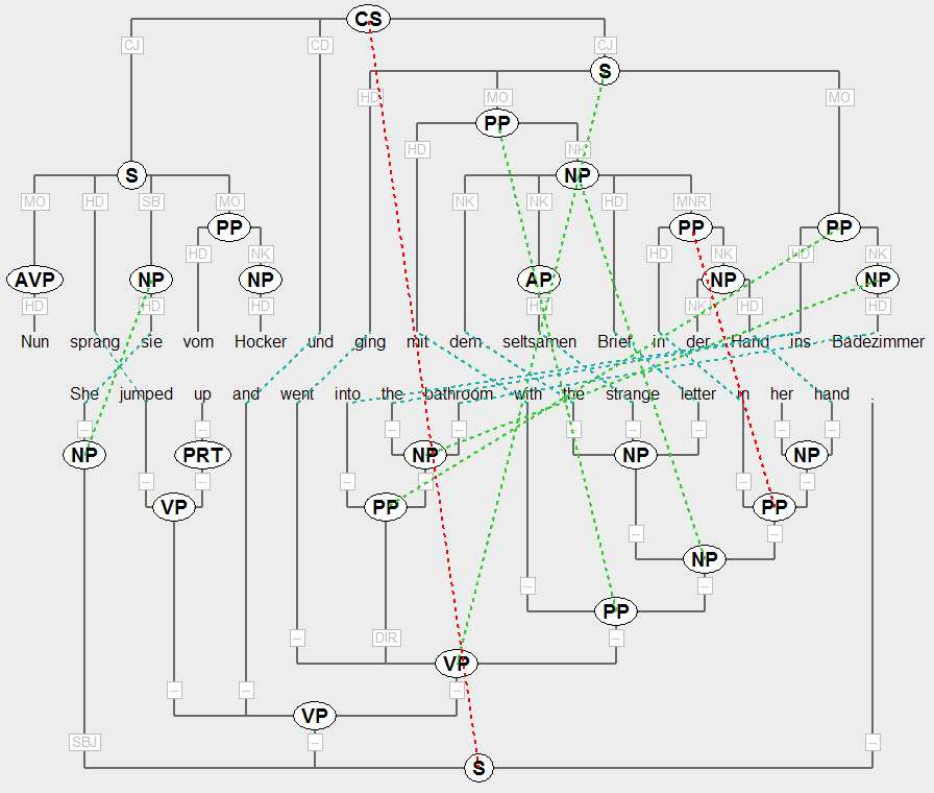

Figure 1: Tree pair German-English with word and phrase alignments.

We do not want to know which part of a sentence has possibly given rise to which part of the correspondence sentence. Instead our perspective is on whether a phrase pair is general enough to be re-used as translation unit in a machine translation system. For example, we do not want to align *die Verwunderung über das Leben* with *their astonishment at the world* although these two phrases were certainly triggered by the same phrase in the original and both have a similar function in the two corresponding sentences. These two phrases seen in isolation are too far apart in meaning to license their re-use. We are looking for correspondences like *was für eine seltsame Welt* and *what an extraordinary world* which would make for a good translation in many other contexts.

Some special rules follow from this principle. For example, we have decided that a pronoun in one language shall never be aligned with a full noun in the other, since such a pair is not directly useful in a machine translation system.

Principles 2 and 3 are more technical. Principle 2 tells our annotators that alignment should be exhaustive. We want to re-use as much as possible from the treebank, so we have to look for as many alignments as possible. And principle 3 says that in case of doubt the alignment should go to the node that is closest to the terminals. For example, our German treebank guidelines require a multi-word proper noun to first be grouped in a PN phrase which is a daughter node of a noun phrase `[[Sofie Amundsen]PN ]NP` whereas the English guidelines only require the NP node `[Sophie Amundsen]NP`. When we align the two names, principle 3 tells us to draw the alignment line between the German PN node and the English NP node since the PN node is closer to the tokens than the German NP node.

Often we are confronted with phrases that are not exact translation correspondences but approximate translation correspondences. Consider the phrases *mehr als eine Maschine* and *more than a piece of hardware*. This pair does not represent the closest possible translation but it represents a possible translation in many contexts. In a way we could classify this pair as the "second-best" translation. To allow for such distinctions we provide our annotators with a choice between exact translation correspondences and approximate correspondences. We also use the term **fuzzy correspondence** to refer to and give an intuitive picture of these approximate correspondences. The option to

distinguish between different alignment strengths sounded very attractive at the start but it turned out to be the source for some headaches later. Where and how can we draw the line between exact and fuzzy translation correspondences?

We have formulated some clear-cut rules:

1. If an acronym is to be aligned with a spelled-out term, it is always an approximate alignment. For example, in our economy reports the English acronym *PT* stands for *Power Technology* and is aligned to the German *Energietechnik* as a fuzzy correspondence.

2. Proper names shall be aligned as exact alignments (even if they are spelled differently across languages; e.g. *Sofie* vs. *Sophie*).

But many open questions persist. Is *einer der ersten Tage im Mai* an exact or rather a fuzzy translation correspondence of *early May*? We decided that it is not an exact correspondence. How shall we handle *zu dieser Jahreszeit* vs. *at this time of the year* where a literal translation would be *in this season*? We decided that the former is still an exact correspondence. These examples illustrate the difficulties that make us wonder how useful the distinction between exact and approximate translation correspondence really is.

Automatically ensuring the overall consistency of the alignment decisions is a difficult task. But we have used a tool to ensure the consistency within the exact and approximate alignment classes. The tool computes the token span for each alignment and checks if the same tokens pairs have always received the same alignment type. For example, if the phrase pair *mit einer blitzschnellen Bewegung* and *with a lightning movement* is once annotated as exact alignment, then it should always be annotated as exact alignment. Figure 1 shows approximate alignments between the PPs *in der Hand* and *in her hand*. It was classified as approximate rather than exact alignment since the German PP lacks the possessive determiner.

Currently our alignment guidelines are 6 pages long with examples for English-German and English-Swedish alignments.

5 Experiments with Student Annotators

In order to check the inter-annotator agreement for the alignment task we performed the following experiment. We gave 20 tree pairs in German and English to 12 advanced undergraduate students in a class on "Machine Translation and Parallel Corpora". Half of the tree pairs were taken from our Sophie's World treebank and the other half from our Economy treebank. We made sure that there was one 1:2 sentence alignment in the sample. The students did not have access to the Gold Standard alignment.

In class we demonstrated the alignment tool to the students and we introduced the general alignment principles to them. Then the students were given a copy of the alignment guidelines. We asked them to do the alignments independently of each other and to the best of their knowledge according to the guidelines.

In our own annotation of the 20 tree pairs (= the Gold Standard alignment) we have the following numbers of alignments:

	type	exact	fuzzy	total
Sophie part	word	75	3	78
	phrase	46	12	58
Economy part	word	159	19	178
	phrase	62	9	71

In the Sophie part of the experiment treebank we have 78 word-to-word alignments and 58 phrase-to-phrase alignments. Note that some phrases consist only of one word and thus the same alignment information is represented twice. We have deliberately kept this redundancy.

The alignments in the Sophie part consist of 125 times 1:1 alignments, 4 times 1:2 alignments and one 1:3 alignment (*wäre* vs. *would have been*) when viewed from the German side. There are 3 times 1:2 alignments (e.g. *introducing* vs. *stellte vor*) and no other 1:many alignment when viewed from the English side.

In the Economy part the picture is similar. The vast majority are 1:1 alignments. There are 207 times 1:1 alignments and 21 times 1:2 alignments (many of which are German compound nouns) when viewed from German. And there are 235 times 1:1 alignments, plus 4 times 1:2 alignments, plus 2 times 1:3 alignments when viewed from English (e.g. the *Americas* was aligned to the three tokens *Nord- und Südamerika*).

The student alignments showed a huge variety in terms of numbers of alignments. In the Sophie part they ranged from 125 alignments to bare 47 alignments (exact alignments and fuzzy alignments taken together). In the Economy part the variation was between 259 and 62 alignments.

On closer inspection we found that the student with the lowest numbers works as a translator and chose to use a very strict criterion of translation equivalence rather than translation correspondence. Three other students at the end of the list are not native speakers of either German and English. We therefore decided to exclude these 4 students from the following comparison.

The student alignments allow for the investigation of a number of interesting questions:

1. How did the students' alignments differ from the Gold Standard?
2. Which were the alignments done by all students?
3. Which were the alignments done by single students only?
4. Which alignments varied most between exact and fuzzy alignment?

When we compared each student's alignments to the Gold Standard alignments, we computed three figures:

1. How often did the student alignment and the Gold Standard alignment overlap?
2. How many Gold Standard alignments did the student miss?
3. How many student alignments were not in the Gold Standard?

The remaining 8 students reached between 81% and 48% overlap with the Gold Standard on the Sophie part, and between 89% and 66% overlap with the Gold Standard on the Economy texts. This can be regarded as their recall values if we assume that the Gold Standard represents the correct alignments. These same 8 students additionally had between 2 and 22 own alignments in the Sophie part and between 12 and 55 own alignments in the Economy part.

So the interesting question is: What kind of alignments have they missed, and which were the additional own alignments that they suggested (alignments that are not in the gold standard)? We first checked the students with the highest numbers of own alignments. We found that some of these alignments were due to the fact that students had ignored the rule to align as close to the tokens as possible (principle 3 above).

Another reason was that students sometimes aligned a word (or some words) with a node. For example, one student had aligned the word *natürlich* to the phrase *of course* instead of to the word sequence *of course*. Our alignment tool allows that, but the alignment guidelines discourage such alignments. There might be exceptional cases where a word-to-phrase alignment is necessary in order to keep valuable information, but in general we try to stick to word-to-word and phrase-to-phrase alignments.

Another discrepancy occurred when the students aligned a German verb group with a single verb form in English (e.g. *ist zurückzuführen* vs. *reflecting*). We have decided to only align the full verb to the full verb (independent of the inflection). This means that we align only *zurückzuführen* to *reflecting* in this example.

The uncertainties on how to deal with different grammatical forms led to the most discrepancies. Shall we align the definite NP *die Umsätze* with the indefinite NP *revenues* since it is much more common to drop the article in an English plural NP than in German? Shall we align a German genitive NP with an of-PP in English (*der beiden Divisionen* vs. *of the two divisions*)? We have decided to give priority to form over function and thus to align the NP *der beiden Divisionen* with the NP *the two divisions*. But of course this choice is debatable.

When we compute the intersection of the alignments done by all students (ignoring the difference between exact and fuzzy alignments), we find that about 50% of the alignments done by the student with the smallest number of alignments is shared by all other students. All of the alignments in the intersection are in our Gold Standard file. This indicates that there is a core of alignments that are obvious and uncontroversial. Most of them are word alignments.

When we compute the union of the alignments done by all students (again ignoring the difference between exact and fuzzy alignments), we find that the number of alignments in the union is 40% to 50% higher than the number of alignments done by the student with the highest number of alignments. It is also about 40% to 50% higher than the number of alignments in the Gold Standard. This means that there is considerable deviation from the Gold Standard.

Comparing the union of the students' alignments to the Gold Standard points to some weak-

nesses of the guidelines. For example, one alignment in the Gold Standard that was missed by all students concerns the alignment of a German pronoun (*wenn **sie** die Hand ausstreckte*) to an empty token in English (*herself __ shaking hands*). Our guidelines recommend to align such cases as fuzzy alignments, but of course it is difficult to determine that the empty token really corresponds to the German word.

Other discrepancies concern cases of differing grammatical forms, e.g. a German definite singular noun phrase (*die Hand*) that was aligned to an English plural noun phrase (*Hands*) in the Gold Standard but missed by all students. Finally there are a few cases where obvious noun phrase correspondences were simply overlooked by all students (*sich - herself*) although the tokens themselves were aligned. Such cases should be handled by an automated process in the alignment tool that projects from aligned tokens to their mother nodes (in particular in cases of single token phrases).

We also investigated how many exact alignments and how many fuzzy alignments the students had used. The following table gives the figures.

	exact	fuzzy	overlap	total
Sophie part	152	106	69	189
Economy part	296	188	119	366

The alignments done by all students resulted in a union set of 189 alignments for the Sophie part and 366 alignments for the Economy part. The alignments in the Sophie part consisted of 152 exact alignments and 106 fuzzy alignments. This means that 69 alignments were marked as both exact and fuzzy. In other words, in 69 cases at least one student has marked an alignment as fuzzy while at least one other student has marked the same alignment as good. So there is still considerable confusion amongst the annotators on how to decide between exact and fuzzy alignments. And in case of doubt many students have decided in favor of fuzzy alignments.

6 Conclusions

We have shown the difficulties in creating cross-language word and phrase alignments. Experiments with a group of students have helped to identify the weaknesses in our alignment guidelines and in our Gold Standard alignment. We have realized that the guidelines need to contain a host of fine-grained alignment rules and examples that will clarify critical cases.

In order to evaluate a set of alignment experiments with groups of annotators it is important to have good visualization tools to present the results. We have worked with Perl scripts for the comparison and with our own TreeAligner tool for the visualization. For example we have used two colors to visualize a student's alignment overlap with the Gold Standard in one color and his own alignments (that are not in the Gold Standard) in another color.

In order to visualize the agreements of the whole group it would be desirable to have the option to increase the alignment line width in proportion to the number of annotators that have chosen a particular alignment link. This would give an intuitive impression of strong alignment links and weak alignment links.

Another option for future extension of this work is an even more elaborate classification of the alignment links. (Hansen-Schirra et al., 2006) have demonstrated how a fine-grained distinction between different alignment types could look like. Annotating such a corpus will be labor-intensive but provide for a wealth of cross-language observations.

References

Ahrenberg, Lars, Magnus Merkel, and Michael Petterstedt. 2003. Interactive word alignment for language engineering. In *Proc. of EACL-2003*, Budapest.

Ahrenberg, Lars. 2007. LinES: An English-Swedish parallel treebank. In *Proc. of Nodalida*, Tartu.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, Genoa.

Groves, Declan, Mary Hearne, and Andy Way. 2004. Robust sub-sentential alignment of phrase-structure trees. In *Proceedings of Coling 2004*, pages 1072–1078, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Hansen-Schirra, Silvia, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the EACL Workshop on Multidimensional Markup in Natural Language Processing (NLPXML-2006)*, pages 35– 42, Trento.

Kruijff-Korbayová, Ivana, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for the Czech-English word alignment. In *Proceedings of LREC*, Genova.

Lundborg, Joakim, Torsten Marek, Maël Mettler, and Martin Volk. 2007. Using the Stockholm TreeAligner. In *Proc. of The 6th Workshop on Treebanks and Linguistic Theories*, Bergen, December.

Melamed, Dan. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-06, IRCS, Philadelphia PA.

Samuelsson, Yvonne and Martin Volk. 2006. Phrase alignment in parallel treebanks. In Hajic, Jan and Joakim Nivre, editors, *Proc. of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 91–102, Prague, December.

Samuelsson, Yvonne and Martin Volk. 2007. Alignment tools for parallel treebanks. In *Proceedings of GLDV Frühjahrstagung 2007*.

Smith, Noah A. and Michael E. Jahr. 2000. Cairo: An alignment visualization tool. In *Proc. of LREC-2000*, Athens.

Tinsley, John, Ventsislav Zhechev, Mary Hearne, and Andy Way. 2007. Robust language pair-independent sub-tree alignment. In *Machine Translation Summit XI Proceedings*, Copenhagen.