Frame-semantic Annotation on a Parallel Treebank

Martin Volk and Yvonne Samuelsson
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
volk@ling.su.se

Abstract

This paper reports on experiments in frame-semantic annotation of a parallel treebank. Selected English and Swedish sentences that contained verbs of motion and communication were annotated independently by two annotators. We found that they assigned the same frame to corresponding sentences in 52% of the cases. This leads us to the conclusion that parallel treebanks can save considerable effort when building semantically annotated resources.

1 The parallel treebank SMULTRON

We have developed a German-English-Swedish parallel treebank, consisting of around 1000 sentences in each language. The first part of our parallel treebank consists of chapters one and two of Jostein Gaarder's novel "Sophie's World". The second part 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.

The name treebank is derived from the fact that syntax structures are mostly encoded as tree graphs. In the annotation we followed the Penn Treebank guidelines for the English trees and the NEGRA/TIGER guidelines for the German trees. For Swedish we adapted the German guidelines. The syntactic annotation for all three languages was done with the ANNOTATE treebank editor. Language-specific chunkers suggested partial trees which were manually checked. This step was followed by automatic tree deepening and extensive consistency checking.

We then aligned the trees in our treebank on the word and phrase level across languages. The alignment is meant to capture translation correspondences in the sense that a phrase pair could be cut out of the trees and reused in an example-based translation system. We distinguish between exact alignment and approximate alignment. This distinction is often debateable but should help if multiple translation alternatives are available for the subsequent MT system. The alignment was done with the TreeAligner, a graphical tool that allows to quickly draw the different alignment lines. We have named our treebank SMULTRON (Stockholm MULtilingual TReebank) and described its development in (Volk and Samuelsson, 2004; Volk et al., 2006), and (Samuelsson and Volk, 2006).

Figure 1 shows an example of parallel trees with word and phrase alignment. The English phrase *"When she crawled through it"* is an exact translation equivalent of *"När hon kröp genom den"* and is therefore aligned with a green line. But the phrase *"a large cavity between the bushes"* is only roughly equivalent to *"en liten håla inne bland buskarna"* (which literally means "a little hole in between the bushes"). Note that we allow m:n sentence alignments and 1:n word and phrase alignments.

The monolingual treebanks are represented in TIGER-XML which defines unique identifiers for all tokens and nodes in the trees. Our alignment uses these identifiers and stores the alignment information in a separate XML file.

2 Frame-semantic Annotation of Parallel Trees

On top of the syntactic annotation we have started to annotate the trees with frame-semantic labels. This was undertaken in student projects for English (Ivantsova, 2006) and for Swedish (Otsa, 2006).[1] In these projects we have focused on frames for motion and communication. 50 trees were handpicked from the Sophie part of our parallel treebank. We made sure that the sentences

[1] Both reports are available at www.ling.su.se/DaLi [Publications].

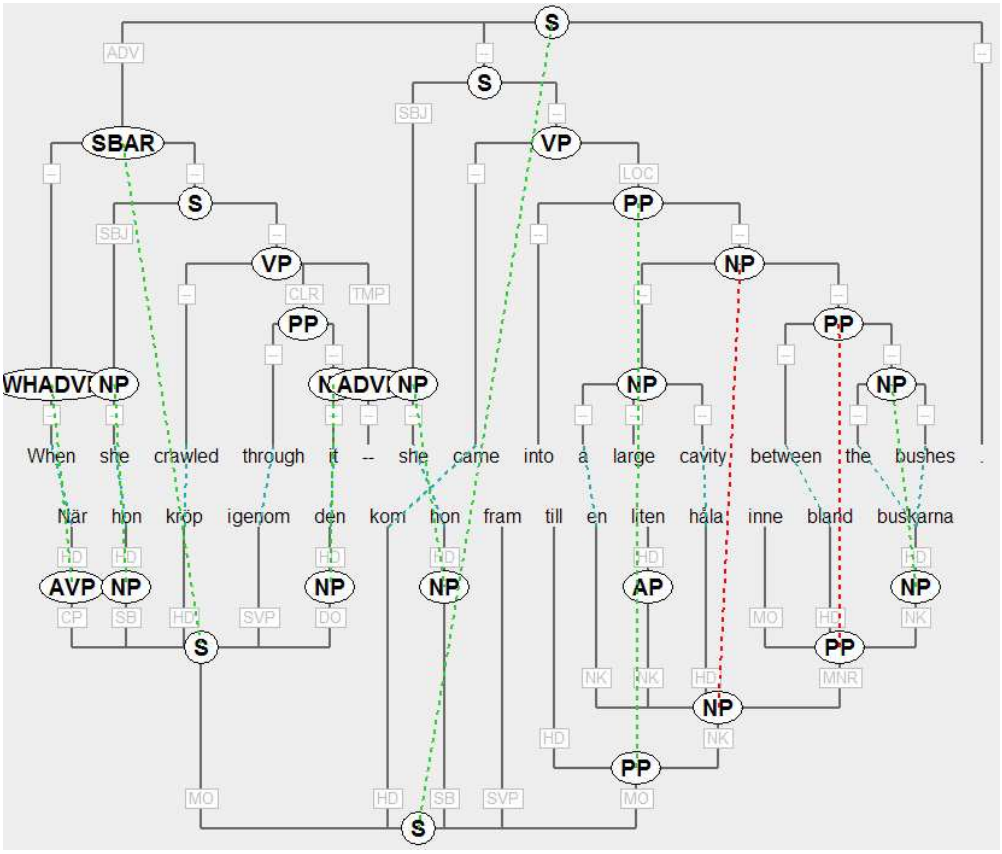

Figure 1: English-Swedish parallel trees with word and phrase alignment.

in both languages contained a verb of motion and communication. For example the English sentence *"She had walked the first part of the way with Joanna"* corresponds to the Swedish sentence *"Den första biten hade hon haft sällskap med Jorunn"*. But while the English sentence contains a motion verb *"walk"*, the Swedish has lost this aspect. It literally translates as *"The first part she had had company with Jorunn"*.

The selected sentences were then independently annotated by the two students in the English and Swedish treebank respectively. The goal of these projects was to see how often the two annotators would assign the same frames in parallel trees. Both used the SALSA tool which was developed for the frame-semantic annotation of German (Erk and Pado, 2004).[2]

Figure 2 shows the result of the frame semantic annotation of the English example tree. It contains the frames Self_motion and Arriving. The Self_motion frame has five elements.[3] The frame elements Self_mover and Path are realized in this sentence and are thus annotated, while Area, Source and Goal are left unattached.

The students used the FrameNet definitions (Fillmore et al., 2003) when they decided which frames and which frame elements to assign. For example, the description of the **Self_motion** frame includes the following definitions:

- The **Self_mover**, a living being, moves under its own power in a directed fashion, i.e. along what could be described as a Path, with no separate vehicle.
- **Goal** is used for any expression which tells where the Self_mover ends up as a result of the motion. E.g. *The children SKIPPED into the park.*
- **Path** is used for any description of a trajectory of motion which is neither a Source nor a Goal. E.g. *The scouts HIKED through the desert.*
- **Source** is used for any expression which implies a definite starting-point of motion. E.g.

[2] See http://www.coli.uni-saarland.de/projects/salsa/

[3] Frame elements are sometimes called "slots" or "roles" in the literature.

The cat RAN out of the house.

- Frame-evoking elements: *crawl, hike, run, skip, walk, ...*

The SALSA tool proved to be very useful for the frame-annotation of both the English and Swedish trees. It takes a TIGER-XML representation of the treebank as its input. It shows a graphical representation of one syntax tree at a time (with or without PoS tags and function labels) and allows the assignment of frames and frame elements. And it saves the result in an extended TIGER-XML file.

The annotator can preselect a set of frames from all defined FrameNet frames. We preselected all frames for motion and communication. The annotator can then assign a frame to a given tree by manually picking from a menu listing. We used eight different motion frames (Arriving, Source_Path_Goal, Body_movement, Cause_motion, Change_direction, Change_posture, Motion, and Self_motion) and six different communication frames (Communication, Communication_noise, Discussion, Questioning, Statement, Telling). Frames were mostly assigned to verbs but sometimes also to phrases (e.g. *was on her way* is assigned a motion frame).

3 Results

For the 50 English sentences 65 frames (17 frame types) were assigned. We list the frames and their frequencies in table 1. The 65 English frames come with 158 instantiated frame elements (26 frame element types). 34 English frames were identical to the frames annotated in the Swedish sentences (52%). In another 22 cases the annotators had assigned closely related frames (e.g. Motion vs. Self_motion) in the two languages. Clear annotation differences arose when the verb choice differed clearly. For example, the English sentence starting with *She was frequently told that ...* in our treebank corresponds to the Swedish *Hon fick ofta höra att ...* (literally: She often got to hear that ...).

This indicates that frame annotation done for one language can be automatically projected to a parallel text. For example, if the semantic frames are annotated for the English sentence "*When she crawled through it she came into a large cavity between the bushes*" (as in figure 2) and when the English syntax tree is aligned to its Swedish counterpart (as in figure 1), then we will be able to automatically transfer the semantic frames to the corresponding Swedish tree. This idea has also been explored by (Pado and Lapata, 2006) for German - English projections on automatic phrase alignments.

Motion	freq
Arriving	2
Body_movement	1
Cause_motion	3
Change_direction	1
Change_posture	1
Cotheme	1
Motion	10
Placing	1
Seeking	1
Self_motion	17
Source-Path-Goal	4
Communication	
Communication	3
Communication_noise	1
Discussion	1
Questioning	6
Statement	9
Telling	3

Table 1: Frames used in the annotation of the English sentences

When we transfer a frame from one sentence to a parallel sentence in another language, then we want both a correct anchoring of the frame in the target language and the correct assignment of the frame elements. This latter step adds to the complexity since some of the frame elements which are realized in the source sentence might not be realized in the target sentence and vice versa.

As a side effect we investigated whether the frames which were originally defined for English were also suitable for Swedish. We found that this was the case. Of course, the selection of the appropriate frames takes more time and effort for Swedish since the Frame-evoking elements (i.e. the verb or phrase triggering a certain set of frames) needs to be translated to English, but then it worked nicely. But we concede that our study was small and therefore we might have missed fine-grained distinctions as found for German-English by (Burchardt et al., 2006). They noticed, for instance, that the "use of dative objects

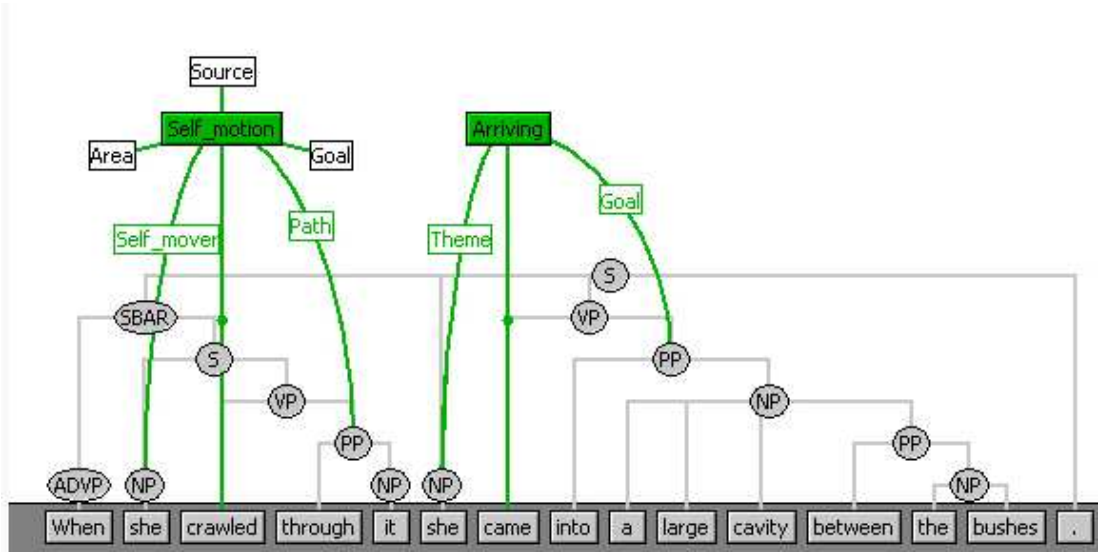

Figure 2: An English syntax tree with frame semantic annotations.

is much less restricted in German than in English". This meant that sometimes an English frame fitted a German sense, but lacked the necessary frame elements. We suspect that similar deviations will eventually arise when porting the English or German frames to Swedish.

4 Conclusions

Our study has demonstrated the usefulness of the SALSA tool and the English frame definitions for frame-semantic annotation of English and Swedish trees. But even more important, it indicates that automatic frame transfer across languages will work in more than 50% of the cases when given a good phrase-alignment. We have not investigated the correctness of the frame element transfer.

Our ultimate goal is to develop a methodology for the large scale annotation and interpretation of parallel texts which is both fast and accurate. Such a methodology will lead to valuable resources for Computational Linguistics, General Linguistics and Translation Studies.

Our parallel treebank provides unique annotation and evaluation material for such a project. We will focus on annotation projection, i.e. to transfer annotation that is computed with certainty for one language to the parallel languages.

References

A. Burchardt, 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.

Katrin Erk and Sebastian Pado. 2004. A powerful and versatile XML format for representing role-semantic annotation. In *Proc. of LREC-2004*, Lisbon.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Natalya Ivantsova. 2006. Enriching a treebank with semantic information in the frame semantics paradigm. C-uppsats, Stockholm University, April.

Annika Otsa. 2006. Berikning av en trädbank med semantisk information. C-uppsats, Stockholms Universitet, April.

Sebastian Pado and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL-COLING 2006*, pages 1161–1168, Sydney, Australia.

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

Martin Volk and Yvonne Samuelsson. 2004. Bootstrapping parallel treebanks. In *Proc. of Workshop on Linguistically Interpreted Corpora (LINC) at COLING*, Geneva.

Martin Volk, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson, and Frida Tidström. 2006. XML-based phrase alignment in parallel treebanks. In *Proc. of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, April.