Swedish-Turkish Parallel Treebank

Beáta Megyesi, Bengt Dahlqvist, Eva Pettersson, Joakim Nivre

Department of Linguistics and Philology
Uppsala University
beata.megyesi@lingfil.uu.se, bengt.dahlqvist@lingfil.uu.se, evapet@stp.lingfil.uu.se, joakim.nivre@lingfil.uu.se

Abstract

In this paper, we describe our work on building a parallel treebank for a less studied and typologically dissimilar language pair, namely Swedish and Turkish. The treebank is a balanced syntactically annotated corpus containing both fiction and technical documents. In total, it consists of approximately 160,000 tokens in Swedish and 145,000 in Turkish. The texts are linguistically annotated using different layers from part of speech tags and morphological features to dependency annotation. Each layer is automatically processed by using basic language resources for the involved languages. The sentences and words are aligned, and partly manually corrected. We create the treebank by reusing and adjusting existing tools for the automatic annotation, alignment, and their correction and visualization. The treebank was developed within the project *Supporting research environment for minor languages* aiming at to create representative language resources for language pairs dissimilar in language structure. Therefore, efforts are put on developing a general method for formatting and annotation procedure, as well as using tools that can be applied to other language pairs easily.

1. Introduction

Parallel corpora containing texts and their translations, where the texts, paragraphs, sentences, and words are typically linked to each other, have become a popular language resource within natural language processing. They have been shown to be useful in several applications, such as machine translation, multi-lingual lexicography, and cross-lingual information retrieval. Parallel corpora are also very useful in empirical language research allowing contrastive studies between languages.

In the past few years, efforts have been made to annotate parallel texts with syntactic structure to build parallel treebanks. A treebank is a syntactically annotated text collection, where the annotation often follows a syntactic theory. Treebanks of today are often based on constituent and/or dependency structure (Abeillé, 2003). A parallel treebank is a parallel corpus where the sentences in each language are syntactically analyzed, and the sentences and words are aligned.

In this paper, we describe a Swedish-Turkish parallel treebank. We build the corpus automatically by using a basic language resource kit (BLARK) for the involved languages and appropriate tools for the automatic alignment and correction of data. The components of the language resource are texts that are in translational relation to each other where the structure is clearly marked up and the sentences and words are analyzed on several linguistic layers.

The treebank we present in this paper is part of the project *Supporting research environment for minor languages* supported by the Swedish Research Council and the Faculty of Languages at Uppsala University. The aim of the project is to create representative language resources for Turkish and Hindi. The Swedish-Turkish treebank serves as a pilot project for building treebanks for other language pairs dissimilar in language structure. Therefore, efforts are put on developing a general method and using tools that can be applied to other language pairs easily.

The paper is organized as follows: section 2 gives an overview of parallel corpora in general and parallel treebanks in particular; section 3 describes the parallel corpus and section 4 presents the method for building the treebank. In section 5, we conclude the paper and suggest some further improvements.

2. Parallell Corpora and Treebanks

In the past years, methods have been developed to build parallel corpora automatically, and to reuse translational data from such corpora for applications. One of the most well-known and frequently used parallel corpora is Europarl (Koehn, 2002) which is a collection of material including 11 European languages taken from the proceedings of the European Parliament.

The largest parallel corpus of today concerning both its size and the number of languages covered is the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006). The corpus consists of documents of legislative text, covering a variety of domains for above 20 languages. Another often used resource is the Bible translated to a large number of languages and collected and annotated by Resnik et al. (1999). The OPUS corpus (Tiedemann and Nygaard, 2004) is another example of a freely available parallel language resource.

There are, of course, many other parallel corpus resources that contain sentences and words aligned in two languages only. Such corpora often exist for languages in Europe, for example the English-Norwegian Parallel Corpus (Oksefjell, 1999) and the ISJ-ELAN Slovene-English Parallel Corpus (Erjavec, 2002).

Parallel treebanks belong to a fairly new type of language resource, consequently we find a smaller amount of resources of this type available. The Prague Czech-English Dependency Treebank (Hajič et al., 2001) is one of the earliest parallel treebanks, containing dependency annotation. The English-German parallel treebank (Cyrus et al., 2003) is another resource with multi-layer linguistic annotation including part of speech, constituent structures, functional relations, and predicate-argument structures. The Linköping English-Swedish Parallel Treebank, also called LinES (Ahrenberg, 2007), currently under development, contains approximately 1,200 sentence pairs,

annotated with part of speech and dependency structures. Stockholm MULtilingual TReebank, also called SMULTRON (Gustafson-Čapková et al., 2007) is a parallel treebank consisting of 1,000 sentences aligned in English, German and Swedish and annotated with constituent structures.
In most parallel corpora including parallel treebanks, we find English and other structurally similar languages. However, there is a need to develop language resources in general, and parallel corpora and treebanks in particular, for other language pairs.
Next, we describe our Swedish-Turkish parallel treebank.

3. Corpus Overview

The corpus, which has been described previously (Megyesi et al., 2006; Megyesi and Dahlqvist, 2007) consists of original texts — both fiction and technical documents — and their translations. The corpus consists of approximately 165,000 tokens in Swedish and 140,000 tokens in Turkish.
After cleaning up the original data received from the publishers, the corpus data is processed automatically by using tools for structural markup, linguistic annotation and sentence and word alignment as shown in figure 1.

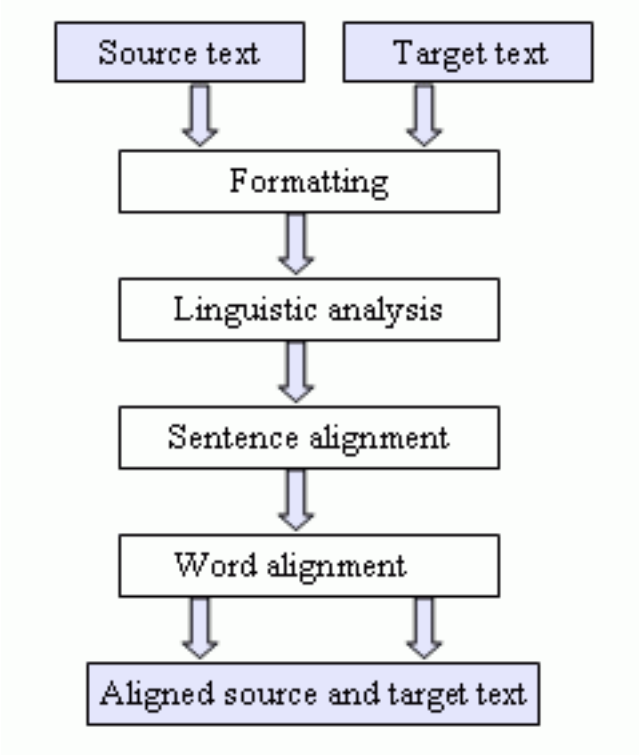

Figure 1: Corpus annotation procedure.

All the essential processing tools are implemented in a graphical user interface (GUI), UplugConnector (Megyesi and Dahlqvist, 2007). The GUI is based on the modules in the Uplug toolkit (Tiedemann, 2003), which consists of a number of perl scripts accessible by line commands with a large number of options and possibility to utilize piping between commands. In addition, various basic language resources developed for the particular languages can be connected to Uplug to be used for text analysis, such as sentence splitting, tokenization, tagging, parsing, and paragraph, sentence and word alignment.
During formatting, the texts are encoded using UTF-8 (Unicode) and marked up structurally using XML Corpus Encoding Standard (XCES) for the annotation format. The plain text files are processed by various tools in the BLARKs of the two languages. The texts are tokenized, the sentences are segmented, the tokens are morphologically analyzed with part of speech and inflectional features. For the linguistic annotation, external morphological analyzers and part of speech taggers are used for the specific languages.
The sentences are aligned automatically, and the words are linked to each other in the two languages. We use standard techniques for the establishment of links between source and target language segments. Paragraphs and sentences are aligned by using the length-based approach developed by Gale and Church (1993). Once the sentences are aligned in the source and target language, we send it for manual correction to a student who speaks both languages. The results show that between 67% and 94% of the sentences were correctly aligned by the automatic aligner depending on the text type.
Phrases and words are aligned using the clue alignment approach (Tiedemann, 2003), and the toolbox for statistical machine translation GIZA++ (Och and Ney, 2003), also implemented in Uplug. Results show that the word aligner aligned approximately 69% of the words correctly.
In addition, we correct the linguistic annotation and alignment manually, and visualize the corpus in different ways without showing the structural markup when used, for example, in teaching.

4. From Parallel Corpus to Parallel Treebank

In order for a parallel corpus to become a parallel treebank, each language in the corpus has to be annotated on the syntactic level. In our treebank, we use several annotation layers for the morpho-syntactic analysis which we describe in this section.
First, we annotate the data morphologically by using external taggers. The Swedish texts are annotated with the Trigrams 'n' Tags tagger (Brants, 2000), trained on Swedish (Megyesi, 2002) with an average accuracy of 96%. The Turkish material is morphologically analyzed and disambiguated using a Turkish analyzer (Oflazer, 1994) and a disambiguator which automatically learns morphological disambiguation rules from a decision list induction algorithm achieving an accuracy of approximately 96% (Yuret and Türe, 2006).
The other linguistic layer contains information about the syntactic analysis. For the grammatical description, we choose dependency rather than constituent structures, as the former has been shown to be well suited for both morphologically rich and free word order languages such as Turkish, and for morphologically simpler languages, like Swedish.
Both the Swedish and the Turkish data were annotated syntactically using MaltParser (Nivre et al., 2006a), trained on the Swedish treebank Talbanken05 (Nivre et al., 2006b)

```
<s id="s20">
    <graph root="p20_4">
      <terminals>
        <t id="w20_1" word="Någon" postag="dt.utr.sin.ind"/>
        <t id="w20_2" word="annan" postag="jj.pos.utr.sin.ind.nom"/>
        <t id="w20_3" word="titel" postag="nn.utr.sin.ind.nom"/>
        <t id="w20_4" word="fanns" postag="vb.prt.sfo"/>
        <t id="w20_5" word="inte" postag="ab"/>
        <t id="w20_6" word="." postag="mad"/>
      </terminals>
      <nonterminals>
        <nt id="p20_1" word="Någon" postag="dt.utr.sin.ind">
          <edge idref="w20_1" label="--"/>
        </nt>
        <nt id="p20_2" word="annan" postag="jj.pos.utr.sin.ind.nom">
          <edge idref="w20_2" label="--"/>
        </nt>
        <nt id="p20_3" word="titel" postag="nn.utr.sin.ind.nom">
          <edge idref="w20_3" label="--"/>
          <edge idref="p20_1" label="DET"/>
          <edge idref="p20_2" label="DET"/>
        </nt>
        <nt id="p20_4" word="fanns" postag="vb.prt.sfo">
          <edge idref="w20_4" label="--"/>
          <edge idref="p20_3" label="SUBJECT"/>
          <edge idref="p20_6" label="PUNC"/>
          <edge idref="p20_5" label="ADV"/>
        </nt>
        <nt id="p20_5" word="inte" postag="ab">
          <edge idref="w20_5" label="--"/>
        </nt>
        <nt id="p20_6" word="." postag="mad">
          <edge idref="w20_6" label="--"/>
        </nt>
      </nonterminals>
    </graph>
  </s>
```

Figure 2: A Swedish sentence represented in Tiger XML.

and on the Metu-SabancıTurkish Treebank (Oflazer et al., 2003), respectively. MaltParser was the best performing parser for both Swedish and Turkish in the CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi, 2006), with a labeled dependency accuracy of 84.6% for Swedish and 65.7% for Turkish. The output from the syntactic parser is in both XCES and Tiger XML. Figure 2 illustrates the representation of the Swedish sentence "Some other title did not exist." as represented in Tiger XML format.

From the Tiger XML format, the syntactic annotation may be visualized with tools like Tiger Search (Lezius, 2002), as illustrated in figure 3 and figure 4, showing dependency graphs for the same sentence "Some other title did not exist", first in Swedish, then in Turkish.

Currently, we manually correct the morphosyntactic annotation in each language as well the output from the word aligner in order to produce some part of the corpus correct thereby making it useful for evaluation.

5. Conclusion

We have presented a Swedish-Turkish parallel treebank — a less processed and typologically dissimilar language pair — containing over 140,000 words in each language. The treebank contains different annotation layers on the morphological and syntactic level using dependency structures. The corpus is automatically created by reusing and adjusting existing tools for the automatic alignment and its visualization, and also partly manually corrected. The Swedish-Turkish parallel treebank is currently used to teach students in Turkish, in linguistic research to study the two languages from a contrastive perspective, and in NLP applications such as for improving word alignment.

In the near future, we are going to use the various linguistic annotations to improve the automatic word alignment, and manually correct the output from the best performing word alignment model(s). In addition, we plan to enlarge the corpus with texts of high translation quality.

6. Acknowledgments

We would like to thank Jörg Tiedemann for his kind support with Uplug, and Kemal Oflazer for the morpho-syntactic annotation of Turkish. The project is financed by the Swedish Research Council and the Faculty of Languages at Uppsala University.

7. References

Anna Abeillé. 2003. *Building and Using Parsed Corpora*. Text, Speech and Language Technology. Kluwer.

Lars Ahrenberg. 2007. Lines: An english-swedish parallel treebank. In *Proceedings of Nordiska Datalingvistdagarna (Nodalida 2007)*.

Thorsten Brants. 2000. Tnt — a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2003. Fuse - a multi-layered parallel treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*.

Tomaz Erjavec. 2002. The ijs-elan slovene-english parallel corpus. *International Journal of Corpus Linguistics*, 7:1:1–20.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Sofia Gustafson-Čapková, Yvonne Samuelsson, and Martin Volk. 2007. SMULTRON (version 1.0) - The Stockholm MULtilingual parallel TReebank. http://www.ling.su.se/dali/research/smultron/index.htm. An English-German-Swedish parallel Treebank with sub-sentential alignments.

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. Prague dependency treebank 1.0 (final production label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0.

Philip Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Technical report, Information Sciences Institute, University of Souther California.

Wolfgang Lezius. 2002. Tigersearch - ein suchwerkzeug für baumbanken (german). In *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*.

Beáta B. Megyesi and Bengt Dahlqvist. 2007. A turkish-swedish parallel corpus and tools for its creation. In *Proceeding of Nordiska Datalingvistdagarna (NoDaLiDa 2007)*.

Beáta B. Megyesi, Anna Sågvall Hein, and Eva Csató Johanson. 2006. Building a swedish-turkish parallel corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

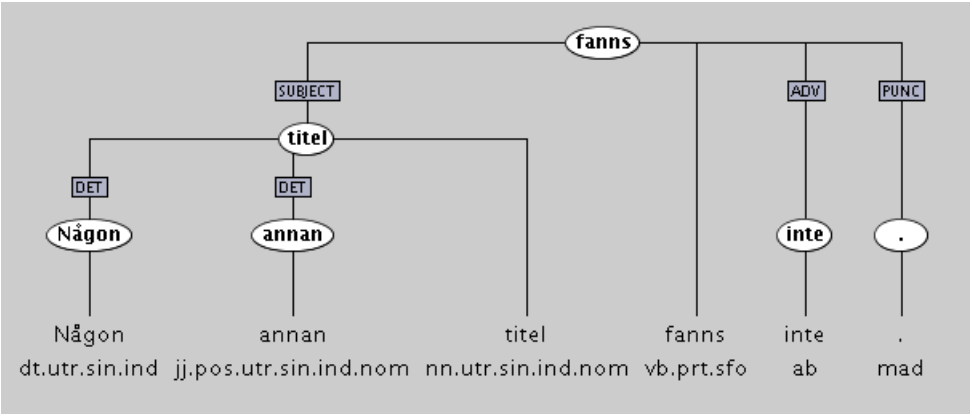

Figure 3: Dependency analysis for the Swedish sentence.

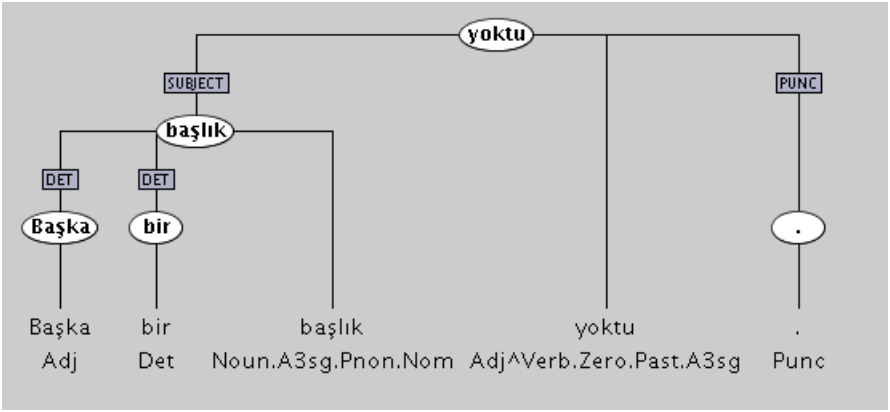

Figure 4: Dependency analysis for the Turkish sentence.

Beáta Megyesi. 2002. *Data-Driven Syntactic Analysis — Methods and Applications for Swedish*. PhD Thesis. KTH.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006b. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:1:19–51.

Kemal Oflazer, Bilge Say, and Dilek Zeynep Hakkani-Tür. 2003. Building a turkish treebank. In *Treebanks: Building and Using Parsed Corpora*.

Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9:2.

Signe Oksefjell. 1999. A description of the english-norwegian parallel corpus: Compilation and further developments. *International Journal of Corpus Linguistics*, 4:2:197–219.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the book of 2000 tongues. *Computers and the Humanitie*, 33(1-2):129–153.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Da'niel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus — parallel & free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Jörg Tiedemann. 2003. *Recycling Translations — Extraction of Lexical Data from Parallel Corpora and their Applications in Natural Language Processing*. PhD Thesis. Uppsala University.

Deniz Yuret and Ferhan Türe. 2006. Learning morphological disambiguation rules for turkish. In *Proceedings of HLT NAACL'06*.