The English-Swedish-Turkish Parallel Treebank

Beáta Megyesi, Bengt Dahlqvist, Éva Á. Csató and Joakim Nivre

Department of Linguistics and Philology, Uppsala University
first.last@lingfil.uu.se

Abstract

We describe a syntactically annotated parallel corpus containing typologically partly different languages, namely English, Swedish and Turkish. The corpus consists of approximately 300 000 tokens in Swedish, 160 000 in Turkish and 150 000 in English, containing both fiction and technical documents. We build the corpus by using the Uplug toolkit for automatic structural markup, such as tokenization and sentence segmentation, as well as sentence and word alignment. In addition, we use basic language resource kits for the linguistic analysis of the languages involved. The annotation is carried on various layers from morphological and part of speech analysis to dependency structures. The tools used for linguistic annotation, e.g., HunPos tagger and MaltParser, are freely available data-driven resources, trained on existing corpora and treebanks for each language. The parallel treebank is used in teaching and linguistic research to study the relationship between the structurally different languages. In order to study the treebank, several tools have been developed for the visualization of the annotation and alignment, allowing search for linguistic patterns.

1. Introduction

Language resources such as linguistically annotated corpora are central components in empirical language studies and natural language processing as they contain authentic language data, which both humans and machines can study and learn from. 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 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 covering a variety of domains for above 20 languages is the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006) consisting of documents of legislative text. 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.

In the past few years, efforts have been made to annotate parallel texts with syntactic structure to build parallel treebanks. A parallel treebank is a parallel corpus where the sentences in each language are syntactically analyzed, and the sentences and words are aligned. In the treebanks, the syntactic annotation usually follows a syntactic theory, often based on constituent and/or dependency structure (Abeillé, 2003). The Prague Czech-English Dependency Treebank (Čmejrek et al., 2004) 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 treebanks, we find English and other structurally similar languages. In the treebank we present in this paper, the user may study structurally dissimilar languages, which also presents challenges for the structural annotation of different language types. The goal of our work is to build a linguistically analyzed, representative language resource for less studied language pairs dissimilar in language structure to be able to study the relations between these languages by researchers, teachers and students.

In this paper, we present a parallel treebank consisting of English, Swedish and Turkish texts. The treebank contains various annotation layers from part-of-speech tags and morphological features to dependency annotation where each layer is automatically annotated, the sentences and words are aligned, and partly manually corrected. We build the corpus automatically using a basic language resource kit (BLARK) for the languages involved and appropriate tools for the automatic alignment and correction of data. The goal is to reuse existing tools as far as possible and develop new ones if necessary for corpus creation, annotation, alignment and visualization.

The work presented in this paper is part of the project *Supporting Research Environment for Less Explored Languages*, supported by the Swedish Research Council and the Faculty of Languages at Uppsala University.

In the next section we describe the data included in the treebank, and in section 3 we give an overview of the method used to create the treebank. In section 4, we give examples of how researchers and students use the resource and in section 5 we conclude the paper and give directions for future research.

2. Treebank Data

The treebank consists of the Swedish-Turkish parallel treebank presented previously (Megyesi et al., 2008) extended with English texts.

The corpus data for each language consists of original texts, both fiction and technical documents, and their translations

Type of Text	English	Swedish	Turkish
The White Castle (O. Pamuk)	-	58 684	44 176
Sofie's world (J. Gaardner)	7 280	7 393	5 651
The royal physician's visit (PO Enquist)	23 323	20 780	16 983
Islam and Europe (I Karlsson)	-	61 529	58 353
Info about Sweden (Migration Office)	-	26 649	28 139
Pregnancy and Giving Birth	1 382	1 076	1 221
Exercise and Food	711	616	685
Psychological Issues	348	385	330
Retirement	-	3 770	4 267
Dublin	496	451	469
UN Declaration of Human Rights	1 911	1 831	1 604
What is unicode	514	539	424
Gospel of Luke	32 238	32 238	-
Gospel of Matthew	29 564	29 247	-
Gospel of Mark	18 872	18 888	-
Gospel of John	24 209	24 625	-
Total	140 848	288 701	162 302

Table 1: Corpus data.

provided by professional translators. The texts vary with respect to translational direction. The majority of the texts is written in Swedish and translated to Turkish and/or English. The treebank contains one novel, The White Castle, written in Turkish and translated to Swedish, and J. Gaardner's novel which is originally written in Norwegian. The corpus also contains UN Declaration of Human Rights, and the four Gospels from the Bible.

In total, the corpus consists of 140,848 tokens in English, 288,701 tokens in Swedish, and 162,302 tokens in Turkish. Table 1 gives an overview of the corpus data with the number of tokens in the three languages. Most of the texts exist in Swedish and Turkish in parallel, but unfortunately we still lack the English translation of the novels and the texts by the Swedish Migration Office which we hope can be included into the treebank in the future.

3. Treebank Development

The texts are processed by various tools developed for each language separately. At the same time, we use the same structural markup and format for all languages. The processing tools are implemented in a framework with a graphical user interface, UplugConnector (Megyesi and Dahlqvist, 2007) which is based on the modules in the Uplug toolkit (Jörg Tiedemann, 2003). Our goal is to produce user-friendly tools to make the annotation, alignment and correction easy for people with less computer skills. The corpus annotation procedure is illustrated in Figure 1.

Independently of language, the original texts are scanned and proof-read, cleaned up and automatically processed. During formatting, the texts are encoded using UTF-8 (Unicode) and marked up structurally using XML Corpus Encoding Standard (XCES) and Tiger XML.

The texts are tokenized, the sentences are segmented, the tokens are morphologically analyzed with part of speech and inflectional features. For the morphosyntactic annotation, external morphological analyzers and part-of-speech taggers are used for the specific languages. The English

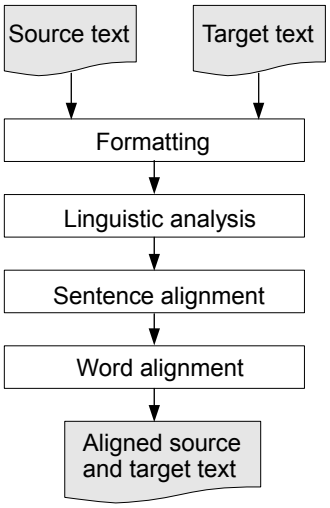

Figure 1: Annotation procedure.

and Swedish texts are annotated with the HunPoS tagger (Halácsy et al., 2007), an open source reimplementation of the Trigrams 'n' Tags tagger (Brants, 2000), with an average accuracy of 96-97% (Megyesi, 2008). The Turkish material is morphologically analyzed 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).

For the syntactic description, we chose 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 English and Swedish.

All data is annotated syntactically using MaltParser (Nivre et al., 2006a), trained on the Penn Treebank for English,

on the Swedish treebank Talbanken05 (Nivre et al., 2006b), and on the Metu-SabancıTurkish Treebank (Oflazer et al., 2003), respectively. MaltParser is one of the best performing dependency parsers for English, Swedish and Turkish, see 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. From the Tiger XML format, the syntactic annotation may be visualized with tools like Tiger Search. Figure 2 illustrates the representation of the Swedish sentence "But he listened attentively." as represented in Tiger XML format.

```
- <s id="s7">
- <graph root="p7_3">
- <terminals>
  <t id="w7_1" word="But" postag="CC" />
  <t id="w7_2" word="he" postag="PRP" />
  <t id="w7_3" word="listened" postag="VBD" />
  <t id="w7_4" word="attentively" postag="RB" />
  <t id="w7_5" word="." postag="." />
 </terminals>
- <nonterminals>
- <nt id="p7_1" word="But" postag="CC">
  <edge idref="w7_1" label="--" />
 </nt>
- <nt id="p7_2" word="he" postag="PRP">
  <edge idref="w7_2" label="--" />
 </nt>
- <nt id="p7_3" word="listened" postag="VBD">
  <edge idref="w7_3" label="--" />
  <edge idref="p7_1" label="VMOD" />
  <edge idref="p7_2" label="SUB" />
  <edge idref="p7_5" label="VMOD" />
  <edge idref="p7_4" label="VMOD" />
 </nt>
- <nt id="p7_4" word="attentively" postag="RB">
  <edge idref="w7_4" label="--" />
 </nt>
- <nt id="p7_5" word="." postag=".">
  <edge idref="w7_5" label="--" />
 </nt>
 </nonterminals>
</graph>
</s>
```

Figure 2: An English sentence "But he listened attentively" represented in Tiger XML.

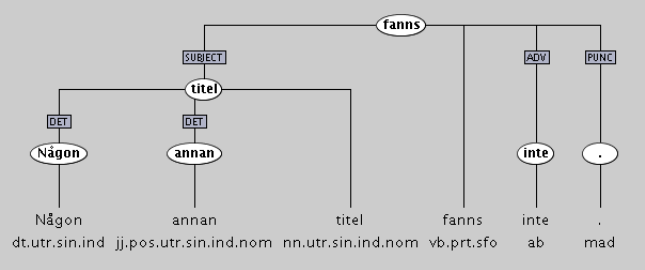

Figure 3: Dependency analysis for the Swedish sentence.

In order to produce a gold standard as part of the corpus, thereby making it useful for training and evaluation, we manually correct the morphosyntactic annotation in each language, focusing on texts for which translations exist for all languages.

After the linguistic analysis, the sentences are aligned automatically, and the words are linked to each other in the language pairs. We use standard techniques for the establishment of links between source and target language segments. Sentences are aligned by using the length-based approach (Gale and Church, 1993). The sentence aligned data is sent for manual correction to a student who speaks both languages. Results show that between 67% and 94% of the sentences were correctly aligned by the automatic aligner depending on the text type in Swedish and Turkish (Megyesi and Dahlqvist, 2007). We calculated the correctness of the sentence alignment results on the first chapter of the novel *White Castle* written by Orhan Pamuk. Not surprisingly, the easiest alignment with 87.3% correctness is the one-to-one mapping between a Swedish and a Turkish sentence. Linking with most errors occurs when several Swedish sentences should have been attached to a single Turkish sentence. The accuracy for the 2-1 alignment between Swedish and Turkish is 33% and for 3-1 is 0%. The 1-0 mapping in the same translation direction also fails in all cases. Evaluation of the sentence alignment results for the other language pairs is in progress.

Words are aligned using the clue alignment approach (Jörg 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 Swedish and Turkish (Megyesi and Dahlqvist, 2007) estimated on 7 000 word pairs in Swedish and Turkish sorted by decreasing frequency taken from the novel *White Castle* written by Orhan Pamuk. The errors in the majority of cases (61%) are due to grammatical differences where multi-word units in Swedish or English often constitute one token in Turkish. We find, for example, unaligned preposition in prepositional phrases in Swedish and English when it should have been linked to the single noun token with a certain case in Turkish.

We are currently extending the treebank with Hindi by including the Uppsala Hindi Corpus consisting of 108,235 tokens (Saxena et al., 2008) to create the Uppsala multilingual treebank. The common part in all four languages at the moment is approximately 5,000 tokens which we hope to be able to extend soon. The low number of joint tokens depends on the lack of texts that are translated to all the languages involved.

4. Applications in Research and Teaching

The treebank is used by researchers, teachers and students in linguistics and Turkish to carry out empirical and contrastive studies. The students can use the corpus directly in their own learning to study various observed linguistic patterns and vocabulary from real texts taken from different genres and different language types. The corpus also serves as a learning platform for testing hypotheses concerning the morphological and syntactic aspects of Turkish

grammar. Further, it helps the students to practice translation between Swedish, English and Turkish. All this is possible due to the fact that the English-Swedish-Turkish parallel texts are available in annotated form. The morpho-syntactic annotations and the alignment are visualized in graphical user interface using pop-up windows.

A search tool has been also developed to help the students to create concordance lists. They can search for whole words, beginnings of words, parts of words or ends of words in all languages. The concordance lists display whole sentences in which the target item appears and it is highlighted. The selected sentences are aligned with their translational equivalents. This form of displaying the linguistic data is considered to be more suitable for learning than KWIC lists in which only the immediate environment of the target item is shown. Figure 4 shows a search result for the Turkish word marked in red color with its morphosyntactic feature in the sentence and its Swedish translation.

5. Conclusion and Future Work

We have presented the English-Swedish-Turkish parallel treebank consisting of over 100,000 words in each language. The treebank contains morphological and syntactic annotation using dependency structures. The corpus is automatically created by reusing and adjusting existing tools for the linguistic analysis, the automatic alignment and its visualization. The corpus is under development and partly manually corrected.

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 manually corrected part of the corpus to be used as gold standard.

6. Acknowledgments

We would like to thank Jörg Tiedemann for his kind support with Uplug, Kemal Oflazer for the morpho-syntactic annotation of Turkish, and Eva Pettersson and Sofia Gustafson-Capkova for their help with the annotation of Swedish and 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*.

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.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos - An Open Source Trigram Tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume Companion Volume, Proceedings of the Demo and Poster Sessions, pages 209–212, Prague, Czech Republic. Association for Computational Linguistics.

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

Philip Koehn. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Technical report, Information Sciences Institute, University of Souther California.

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, Bengt Dahlqvist, Eva Pettersson, and Joakim Nivre. 2008. Swedish-Turkish Parallel Treebank. In *Proceeding of Language Resources and Evaluation (LREC 2008)*.

Beáta Megyesi. 2008. The Open Source Tagger HunPos for Swedish. In *Report, Department of linguistics and philology, Uppsala University*.

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.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the Book

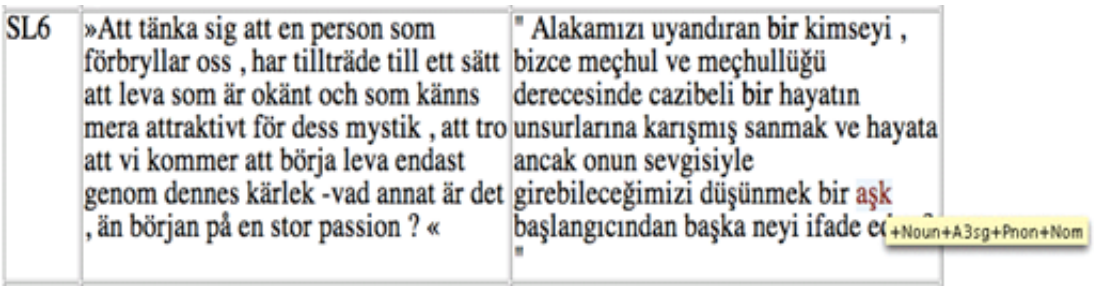

Figure 4: Example taken from the visualization tool.

of 2000 Tongues. *Computers and the Humanities*, 33(1-2):129–153.

Anju Saxena, Pranava Swaroop Madhyasta, and Joakim Nivre. 2008. Building the Uppsala Hindi Corpus. In *Proceedings of the second Swedish Language Technology Conference*.

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)*.

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. In *4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Deniz Yuret and Ferhan Türe. 2006. Learning Morphological Disambiguation Rules for Turkish. In *Proceedings of HLT NAACL'06*.