Supporting Research Environment for Swedish and Turkish

Beáta B. Megyesi
Éva Csató Johanson
Bengt Dahlqvist
Sofía Gustafson-Capková
Joakim Nivre
Eva Pettersson
Anna Sågvall Hein

Department of Linguistics and Philology
Uppsala University

1 Introduction

Language resources such as corpora consisting of annotated texts and utterances have been shown to be a central component in language studies and natural language processing as they, when carefully collected and compiled, contain authentic language material capturing information about the language. Corpora are shown to be useful in language research allowing empirical studies, as well as for various applications in natural language processing. During the last decade, researchers’ attention have been directed to building parallel corpora including texts and their translations as they contain highly valuable linguistic data across languages. Methods have been developed to build parallel corpora by automatic means, and to reuse translational data from such corpora for several applications, such as machine translation, multi-lingual lexicography and cross-lingual domain-specific terminology. Parallel corpora exist for many language pairs, mainly European languages with special focus on Western-Europe.

In the past few years, efforts have been made to annotate parallel texts on different linguistic levels up to syntactic structure to build parallel treebanks. A treebank is a syntactically annotated text collection, where the annotation often follows a syntactic theory, mainly 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.

The primary 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. The aim is to build a parallel treebank containing 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. The work described here is part of the project Supporting research environment for minor languages initiated by professor Anna Sågvall Hein at Uppsala University. The project aims at building various types of language resources for Turkish and Hindi. We choose Swedish and Turkish, a less studied and typologically dissimilar language pair, to serve as a pilot study for building parallel treebanks for other language pairs. Therefore, efforts are put on developing a general method and using tools that can be applied to other language pairs easily.

The components of the language resource are texts that are in translational relation to each other and syntactically analyzed, and tools for the automatic analysis and alignment of these languages. To build a parallel corpus, we reuse existing resources and create necessary tools for the automatic processing and alignment of the parallel texts in these languages. The purpose is to build the corpus automatically by using a basic language resource kit (BLARK) for the particular languages and appropriate tools for the automatic alignment and correction of data. We use tools that are user-friendly, understandable and easy to learn by people with less computer skills, thereby allowing researchers and students to align and correct the corpus data by themselves. The parallel treebank is intended to be used in linguistic research, teaching and applications such as machine translation.

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 languages and the data the treebank contains, the methods used for building the treebank and the tools used for visualization, correction and investigation of the treebank. In section 4, we conclude the paper.
2 Parallel Corpora and Parallel Treebanks

A parallel corpus is usually defined as a collection of original texts translated to another language where the texts, paragraphs, sentences, and words are typically linked to each other. 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. Another parallel corpus is the JRC-Acquis Multilingual Parallel Corpus (Steinberger et al., 2006). It is the largest existing parallel corpus of today concerning both its size and the number of languages covered. 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). It is especially common to include English as one of the two languages in the pair. Parallel corpora that do not include English or another European language are rare.

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 (Hajic 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. There are also small parallel treebanks including Swedish as one of the languages under development. The Linköping English-Swedish Parallel Treebank, also called LinES (Ahrenberg, 2007) contains approximately 1200 sentence pairs, annotated with PoS and dependency structures, and the Swedish-English-German treebank, SMULTRON (Gustafson-Capkova et al., 2007), annotated with PoS and 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 the development of our Swedish-Turkish parallel treebank.

3 The Swedish-Turkish Parallel Treebank

Before we present the content and the annotation procedure of the treebank, we give a short overview of the involved languages as they are less known.

3.1 A Note on Swedish and Turkish

Swedish belongs to the Scandinavian, North Germanic family of the Germanic branch of the Indo-European languages. It is an inflective language and morphologically richer than for example English. Nouns in general have a two gender distinction and are marked by articles, adjectives, anaphoric pronouns. As in English, nouns can appear with or without articles. There are definite and indefinite articles that agree with the head noun in gender, number and definiteness. Furthermore, adjectives have gender, definiteness and plural markers. Also, compound nouns composed as single words are frequent and productive. Verbs lack markers for person or number of the subject but retain tense including complex tense forms. From a syntactic point of view, Swedish has subject-verb-object (SVO) order in independent declarative sentences, as well as in subordinate clauses, similar to English. However, in subordinate clauses the sentence adverbs normally precede the finite verb and the perfect auxiliary can be omitted.

Turkish is not an Indo-European language. It is a Turkic language and belongs to the Altaic branch of the Ural-Altaic family. It is a suffixing and agglutinative language; in most of the cases, there is a one-to-one relationship between morpheme and function. The vowels of suffixes undergo vowel harmony with respect to backness and rounding. There are five cases: genitive, dative, accusative, locative, and ablative. The verbal system is rich and verbs have markers for tense, mood, aspect, and voice, as well as agreement markers in terms of the features person and number. Considering the syntactic characteristics, Turkish is a left-branching type of language, where the dependents precede their head (for example adjective or genitive modifier precedes the modified head, and objects precede the verb). Turkish is rather free in its word order which is based on the information structure. The unmarked word order is SOV (verb final word order) but
other orders are possible depending on which element is put into the focus in the discourse. Subordinate clauses are often constructed by infinite constructions. Turkish is a pro-drop language, that is subjects can be left unexpressed in finite clauses because of the rich agreement morphology.

3.2 Corpus Content

The corpus, which has been previously described (Megyesi et al., 2006; Megyesi & Dahlqvist, 2007; and Megyesi et al., 2008) consists of original texts – both fiction and technical documents – and their translations from Turkish to Swedish and from Swedish to Turkish with the exception of one text which is a translation from Norwegian to both languages. In table 1, the corpus material is summarized.

The corpus consists of approximately 165,000 tokens in Swedish and 140,000 tokens in Turkish. Divided into text types, the fiction part of the corpus includes 76,877 tokens in Swedish, and 55,378 tokens in Turkish. The technical documents are larger and contain 90,901 tokens in Swedish, and 85,171 tokens in Turkish. The current material presented here serves as pilot linguistic data for the Swedish-Turkish parallel corpus. We intend to extend the material to other texts, both technical and fiction, in the future.

<table>
<thead>
<tr>
<th>Document: Fiction</th>
<th># Tokens</th>
<th># Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>The White Castle — Swedish</td>
<td>53232</td>
<td>7748</td>
</tr>
<tr>
<td>The White Castle — Turkish</td>
<td>36684</td>
<td>12472</td>
</tr>
<tr>
<td>Sofie's world — Swedish</td>
<td>6488</td>
<td>1466</td>
</tr>
<tr>
<td>Sofie's world — Turkish</td>
<td>4800</td>
<td>2215</td>
</tr>
<tr>
<td>The Royal Physician's Visit — Swedish</td>
<td>17157</td>
<td>3932</td>
</tr>
<tr>
<td>The Royal Physician's Visit — Turkish</td>
<td>13894</td>
<td>5456</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document: Non-fiction</th>
<th># Tokens</th>
<th># Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Islam and Europe — Swedish</td>
<td>55945</td>
<td>10977</td>
</tr>
<tr>
<td>Islam and Europe — Turkish</td>
<td>48893</td>
<td>14128</td>
</tr>
<tr>
<td>Info about Sweden — Swedish</td>
<td>24107</td>
<td>4576</td>
</tr>
<tr>
<td>Info about Sweden — Turkish</td>
<td>23660</td>
<td>7119</td>
</tr>
<tr>
<td>Retirement — Swedish</td>
<td>3417</td>
<td>818</td>
</tr>
<tr>
<td>Retirement — Turkish</td>
<td>3664</td>
<td>1188</td>
</tr>
<tr>
<td>Dublin — Swedish</td>
<td>392</td>
<td>169</td>
</tr>
<tr>
<td>Dublin — Turkish</td>
<td>394</td>
<td>230</td>
</tr>
<tr>
<td>Pregnancy — Swedish</td>
<td>949</td>
<td>409</td>
</tr>
<tr>
<td>Pregnancy — Turkish</td>
<td>1042</td>
<td>567</td>
</tr>
<tr>
<td>Psychology — Swedish</td>
<td>347</td>
<td>193</td>
</tr>
<tr>
<td>Psychology — Turkish</td>
<td>281</td>
<td>220</td>
</tr>
<tr>
<td>Movement — Swedish</td>
<td>543</td>
<td>300</td>
</tr>
<tr>
<td>Movement — Turkish</td>
<td>568</td>
<td>369</td>
</tr>
<tr>
<td>Social security — Swedish</td>
<td>5201</td>
<td>846</td>
</tr>
<tr>
<td>Social security — Turkish</td>
<td>6669</td>
<td>2025</td>
</tr>
</tbody>
</table>

Table 1: The corpus data divided into text categories with number of tokens and types.
3.3 Corpus Annotation

The corpus material is processed automatically by using various tools making the annotation, alignment and manual correction easy and straightforward for users with less computer skills. This is necessary, as our ambition is to allow researchers and students of particular languages to enlarge the corpus by automatically processing and correcting the new data by themselves.

First, the original materials, i.e., the source and target texts received from the publishers in various formats are cleaned up. For example, rtf, doc, and pdf documents are converted to plain text files. In the case of the original pdf-file, we scan and proof-read the material and, where necessary, correct it to ensure that the plain text file is complete and correct. After cleaning up the original data, the texts are processed automatically by using tools for formatting, linguistic annotation and sentence and word alignment. Figure 1 gives an overview of the main modules in the corpus annotation procedure as processed in Uplug.

![Diagram of corpus annotation process]

Figure 1: The modules of corpus annotation.

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 then processed by various tools in the BLARKs developed for each language separately when necessary. A sentence splitter is used to break the texts into sentences, and a tokenizer is used to separate words from punctuation marks. An example of the output from the formatter taken from Orhan Pamuk’s book *The White Castle* is shown for the sentence “Some other title did not exist.” in example 1 in Swedish, and in example 2 in Turkish.

```xml
<s id="s11.4">
  <w id="w11.4.1">Någon</w>
  <w id="w11.4.2">annan</w>
  <w id="w11.4.3">titel</w>
  <w id="w11.4.4">fanns</w>
  <w id="w11.4.5">inte</w>
</s>
```

Example 1: XML-annotation of the Swedish sentence “Någon annan titel fanns inte”.
Example 2: XML-annotation of the Turkish sentence “Başka bir başlık yoktu.”

Once the sentences and tokens are identified, the data is linguistically analyzed. For the linguistic annotation, external morphological analyzers, part-of-speech taggers and syntactic dependency parsers are used for the specific languages. We use several annotation layers for the linguistic analysis, first on a morphological level, then on a syntactic level.

The Swedish texts are morphologically annotated with the open source HunPoS tagger (Halacy, et al., 2007), which is an implementation of the Trigrams’ n’ Tags part of speech tagger (Brants, 2000), trained on Swedish (Megyesi, 2008) using the Stockholm-Umeå Corpus (SUC, 1997). The tokens are annotated with parts of speech and morphological features and are disambiguated according to the syntactic context. In example 3, the morphological annotation in XCES annotation format is shown for the same Swedish sentence as previously.

Example 3: XCES-annotation with morphological information of the Swedish sentence “Någon annan titel fanns inte.”.

The results for the morphological annotation of Swedish show an accuracy of 96.6%. The most erroneous tags in the materials are: i) proper nouns which should be tagged as common nouns, ii) particles which should be tagged as adverbs, iii) prepositions which should be annotated as particles or adverbs, iv) nouns with wrong morphological analysis and finally v) participles which should be tagged as verbs. These errors constitute 46% of all errors.

The Turkish material is analyzed morphologically by using an automatic morphological analyzer developed for Turkish (Oflazer, 1994). Each token in the text is segmented and annotated with morphological features including part-of-speech. The Turkish material is morphologically analyzed and disambiguated using a Turkish analyzer (Oflazer, 1994) and a disambiguator (Yuret & Türe, 2006). Example 4 shows the morphological annotation in XCES annotation format for the Turkish sentence.

Example 4: XCES-annotation with morphological information of the Turkish sentence “Başka bir başlık yoktu.”

Evaluation of the Turkish tagging and disambiguation shows an average accuracy of 78.6%. Problematic confusions in the Turkish tagging seems to be between i) determiners and numerals, ii) postpositions in nominative and postpositions in genitive, and iii) determiners and pronouns. These errors stand for 24.9% of all errors.
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 are annotated syntactically using MaltParser (Nivre et al., 2006a), trained on the Swedish treebank Talbanken05 (Nivre et al., 2006b) and on the Metu-Sabancı Turkish 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. Example 5 and 6 shows the morphological and syntactic annotation for the same sentence as previously, formatted as XCES.

Example 5: XCES-annotation with morphological and syntactic information of the Swedish sentence “Någon annan titel fanns inte.”

Example 6: XCES-annotation with morphological and syntactic information of the Turkish sentence “Başka bir başlık yoktu.”

In addition, an equivalent annotation is also produced in TIGER XML format in order to use existing tools for visualization and correction of the parallel treebank, as well as searching in trees. We will return to the tools in the next chapter. The representation in TIGER XML for the same sentence is shown in example 7 and 8, first in Swedish, then in Turkish.

Example 5: XCES-annotation with morphological and syntactic information of the Swedish sentence “Någon annan titel fanns inte.”

Example 6: XCES-annotation with morphological and syntactic information of the Turkish sentence “Başka bir başlık yoktu.”

In addition, an equivalent annotation is also produced in TIGER XML format in order to use existing tools for visualization and correction of the parallel treebank, as well as searching in trees. We will return to the tools in the next chapter. The representation in TIGER XML for the same sentence is shown in example 7 and 8, first in Swedish, then in Turkish.
Example 7: TIGER-XML annotation including linguistic information of the Swedish sentence “Någon annan titel fanns inte.”.
Example 8: TIGER-XML annotation including linguistic information of the Turkish sentence “Başka bir başlık yoktu.”

After the linguistic analysis, 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). The aligned sentences are stored in XML format, as shown in example 9 below.

```xml
<cesAlign toDoc="vt.xml" version="1.0" from Doc="vs_tnt.xml">
  <linkGrp targType="s" toDoc="vt.xml" from Doc="vs_tnt.xml">
    <link certainty="8" xtargets="s1.1;s1.1" id="SL0.1"/>
    <link certainty="111" xtargets="s2.1;s2.1" id="SL0.2"/>
    <link certainty="-1287" xtargets="s3.1;s2.2 s3.1" id="SL0.3"/>
    <link certainty="-340" xtargets="s3.2;s3.2" id="SL0.4"/>
    <link certainty="114" xtargets="s3.3;s3.3" id="SL0.5"/>
  ...
</cesAlign>
```

Example 9: XCES-annotation of the sentence alignment.

Once the sentences are aligned in the source and target language, we send it for manual correction to a student who speaks both languages. We automatically compare the links before and after the manual correction and the user gets statistics about the differences. The results show that between 67% and 94% of the sentences were correctly aligned by the automatic sentence aligner depending on the text type.

The distribution of the alignment types after the manual alignment is shown in column two in table 2 for the first chapter of The White Castle. The manually corrected alignment resulted in 178 sentence pairs after merges and splits. The automatic sentence alignment produced 168 sentence pairs. The correctness of these compared with the manual alignment is presented in column three and four. Our results show that 74.4% of the sentences were correctly aligned by the automatic aligner. All one-to-two links and 87.3% of the one-to-one mappings are correct. The lowest score are the two-to-one alignments, where 33% are correctly aligned.

<table>
<thead>
<tr>
<th>Link type: Swedish-Turkish</th>
<th>Manual Number</th>
<th>Automatic Number</th>
<th>Correct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1-1</td>
<td>144</td>
<td>126</td>
<td>110 (87.3)</td>
</tr>
<tr>
<td>1-2</td>
<td>3</td>
<td>3</td>
<td>3 (100)</td>
</tr>
<tr>
<td>2-1</td>
<td>15</td>
<td>39</td>
<td>12 (33.0)</td>
</tr>
<tr>
<td>3-1</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>178</strong></td>
<td><strong>168</strong></td>
<td><strong>125 (74.4)</strong></td>
</tr>
</tbody>
</table>

Table 2: Distribution of manual alignment for various link types and the result of the automatic sentence alignment.

Lastly, phrases and words are aligned using the clue alignment approach (Tiedemann, 2003), and the toolbox for statistical machine translation GIZA++ (Och and Ney, 2003). Results show that the word aligner aligned approximately 69% of the words correctly. For a detailed evaluation of the results, we investigated the error level on 7,077 word pairs in Swedish and Turkish sorted by decreasing frequency taken from The White Castle. Of the incorrectly aligned pairs that appeared at least twice in the material, 61% of the errors can be considered due to grammatical
differences between the two languages. Often, Swedish has an expression of several tokens while Turkish expresses the same in one token. For example, the aligner often fails to attach the preposition (till, 'to') in prepositional phrases in Swedish (till sultanen, 'to the sultan') to the single Turkish word (padişahın). The aligner also fails to attach the subordinate conjunction (som, 'that') and the 3rd person pronoun (han, 'he') in the Swedish utterance (som han ville, 'that which he wanted') to the Turkish segment expressed as one single word, the verb (istediğini, 'that what he wanted') since Turkish is a pro-drop language and can leave out the pronominal subject and the relative clause is constructed as various participial forms as verbal suffixes. The remaining errors, which constitute approximately 39% of the wrongly aligned material, cannot be explained by grammatical differences between the two languages. Rather, these might appear as a consequence of the previously occurring errors in the alignment.

In addition to the automatic morpho-syntactic annotation and alignment, we correct the linguistic analysis and links manually, and visualize the corpus in different ways without showing the structural markup when used, for example, in teaching. These tools will be described next.

### 3.4 Tools for visualization and correction

In the project, our goal is to reuse and further develop freely available, system independent, user-friendly tools for the annotation, visualization, correction and search in our corpus, both considering the mono-lingual and the parallel treebanks.

As basis for the annotation, we use the Uplug toolkit which is a collection of tools for processing corpus data, created by Jörg Tiedemann (2003). Uplug is used for sentence splitting, tokenization, tagging by using external taggers, and paragraph, sentence and word alignment. All the essential processing tools are implemented in a graphical interface, UplugConnector (Megyesi and Dahqvist, 2007) which accesses both the modules in the Uplug toolkit (Tiedemann, 2003), and other programs for linguistic annotation.

The Uplug package consists of a number of perl scripts accessible by line commands with a large number of options and sometimes utilizing piping between commands. To facilitate easier access and usage of these scripts, a graphical user interface, UplugConnector, was developed in Java for the project. Here, the user can in a simple fashion choose a specific task to be performed and let the graphical user interface (GUI) set up the proper sequence of calls to Uplug and subsequently execute them. Figure 2 below illustrates the Uplug Connector interface.

![Figure 2: The Uplug Connector](image)

The user can optionally give the location of the source and target files, decide where the output should be saved, and specify the encoding for the input and output files. For the markup, basic structural markup, sentence segmentation, and tokenization are available. Further, the Uplug Connector GUI has been constructed to give the possibility to include calls to new scripts outside Uplug for complementary analysis, when such needs arise. The user can easily access another resource if the available ones do not fit his/her needs, for example an external tokenizer, sentence splitter, tagger or parser. In the toolkit, the user can also call for the sentence and word aligners and their visualization tools.
As the XML representation of the result is not user friendly even for people used to this kind of annotation, we use various interfaces for the visualization of the linguistic annotation and alignment results. In addition, since the automatic alignment generates some errors, we also use tools for the manual correction of these.

As a tool for the correction of the sentence alignment, we choose the system ISA (Interactive Sentence Alignment) developed by Tiedemann (2006). ISA is a graphical interface for automatic and manual sentence alignment which uses the alignment tools implemented in Uplug. It handles the manual correction of the sentence alignment in a user-friendly, interactive fashion. Figure 3 shows ISA with the aligned sentences taken from Orhan Pamuk's book *The White Castle*.

For displaying the corrected sentence output from ISA after manual correction of the alignment together with the linguistic analysis, a script utilizing the structural XML-parser Hpricot (2006) was developed. It takes as input the tagged XML-files for the language pair together with the XML file containing the sentence alignment results produced by ISA and generates an HTML-file which displays the sentences aligned together with the morphological information for each word shown in pop-up windows.

The visualization tool makes it easier for students and researchers to study the part of speech and inflectional features of the words and chosen structures for translation than the structurally marked up version of the corpus.

![ ISA showing the aligned sentences from The White Castle. ]

![ Visualization of aligned sentence pairs with linguistic annotation shown in the pop-up window. ]
To visualize the word alignment result in a simple way, a new script for HTML-visualization of the word alignment result was included in the Uplug Connector. This takes as input the text file with word link information produced by Uplug, see figure 5, and shows the word-pair frequencies. This visualization actually presents a bilingual lexicon created from the source and target language data.

![Figure 5: HTML-visualization of word alignment.](image)

To manually correct the automatically generated linguistic features, such as the part-of-speech tags in the tagger output XML-files, a web-based tool POS-CORR was developed. The web form is easier, safer and faster to use than making the corrections in an ordinary text editor, see figure 6. The tool controls that only predefined tag constituents (for the moment are 107 included) can be entered, and produces a new XML-file with the correct corrections inserted.

![Figure 6: POS-CORR, a tool for manual correction of morphological features.](image)

For the visualization and correction of the parallel syntactic trees, several freely available tools have been investigated. I-Link (Ahrenberg, Merkel & Andersson, 2002) is a Java application originally designed for interactive manual correction of output from the automatic word alignment tool Linköping Word Aligner. The user works in a graphical interface, where the source and target sentence(s) are displayed side by side, and words that are aligned have the same color. The user may easily add or delete links by clicking the words that should be selected for the operation. The system takes as input XML-annotated source and target files and a set of token links. Alpaco (Aligner for Parallel Corpora) (Rassier and Pedersen, 2003) is a Perl Tk application, originally designed for manually aligning parallel texts at word-by-word or phrase-by-phrase level. Input is raw text files or previously aligned text files in Blinker or Alpaco format. UMIACS Word Alignment Interface (Hwa and Madnani, 2004) is a Java tool where the source

1 See [http://www.d.umn.edu/~tpederse/Code/Readme.Alpaco-v0.3.txt](http://www.d.umn.edu/~tpederse/Code/Readme.Alpaco-v0.3.txt)
2 See [http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm](http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm)
and target sentences are displayed side by side in a graphical interface, in a similar way as in I-Link, and the user can add links by clicking the words to be selected.

Since we are interested in visualizing, inspecting and correcting tree structures, we finally chose Stockholm Tree Aligner (Lundborg, et al., 2007) to be used in the project. The tool also allows the user to create links between corresponding nodes in two treebanks, hence allowing word and phrase alignment correction between our languages. The input to Stockholm Tree Aligner takes trees in TigerXML format showing edge labels and crossing branches. The tool also contains a search function that implements the TigerSearch Query Language with additions for searching alignments. The sentence “Some other title did not exist.” is visualized as syntactic trees for Turkish and Swedish showing the dependency relations between the elements in each sentence in figure 7.

![Diagram of dependency relations in Turkish and Swedish](image)

**Figure 7:** Dependency relations in Turkish and Swedish for the same sentence.

In addition to the visualization of the linguistic annotation and alignments, the user can also search for word forms in a subset of the corpus in both languages through a graphical interface given as a web page. Figure 8 illustrates the interface where the user can choose a text for searching, give a word form in a specific language (Turkish or Swedish), and search for the whole word, or the beginning, the end or a part of the word. The data is stored in an sql-database with the sentences as the single retrievable entries. The found matching sentences in which the word form appears together with the aligned sentences in the other language are displayed in a result screen. Information about the text name, the word form, the number of matching sentences, and word strings is also shown, as illustrated in figure 8. The interface has been of particular use by the students in their work, and it serves as a complement to the previous full corpora display facilities developed in the project.

---

Figure 8: Search function for word forms through a graphical interface as a web page.

Figure 9: Search result for a word form in a particular text together with the number of matching sentences and word strings.
4 Conclusion

We have presented a Swedish-Turkish parallel treebank – a less processed language pair – containing approximately 150,000 tokens in Swedish, and 126,000 tokens in Turkish. The treebank is automatically created by re-using and adjusting existing tools for the automatic alignment and its visualization, and basic language resource kits for the automatic linguistic annotation of the involved languages. The automatic annotation and alignment is also partly manually corrected. The treebank is already in use in language teaching, primarily in Turkish.

Acknowledgments

We are grateful to Jörg Tiedemann for his kind support with Uplug, and Kemal Oflazer and Gülşen Eryigit for the morpho-syntactic annotation of the Turkish material. Also, we would like to thank the publishers for allowing us to use the texts in the corpus. The project is financed by the Swedish Research Council and the Faculty of Languages at Uppsala University.

References


92-101, Prague.


