

# Supporting Research Environment for Less Explored Languages: A Case Study of Swedish and Turkish

Beáta Megyesi  
Bengt Dahlqvist  
Eva Pettersson  
Sofia Gustafson-Capková  
Joakim Nivre

Uppsala University  
Department of Linguistics and Philology

## 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 Sagvall 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 parallel treebank, the methods used for building the treebank and the tools used for visualization, correction and investigation of the treebank. In section 4, we suggest some further improvements and lastly, in section 5, 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-Capková 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**

First, we present the content and the annotation procedure of the treebank, then we give an overview of the tools that we use for the visualization and correction of the corpus annotation.

#### **3.1 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.

<i>Document: Fiction</i>	<i># Tokens</i>	<i># Types</i>
The White Castle – Swedish	53232	7748
The White Castle – Turkish	36684	12472
Sofie’s world – Swedish	6488	1466
Sofie’s world – Turkish	4800	2215
The Royal Physician’s Visit – Swedish	17157	3932
The Royal Physician’s Visit – Turkish	13894	5456
<i>Document: Non-fiction</i>		
Islam and Europe – Swedish	55945	10977
Islam and Europe – Turkish	48893	14128
Info about Sweden – Swedish	24107	4576
Info about Sweden – Turkish	23660	7119
Retirement – Swedish	3417	818
Retirement – Turkish	3664	1188
Dublin – Swedish	392	169
Dublin – Turkish	394	230
Pregnancy – Swedish	949	409
Pregnancy – Turkish	1042	567
Psychology – Swedish	347	193
Psychology – Turkish	281	220
Movement – Swedish	543	300
Movement – Turkish	568	369
Social security – Swedish	5201	846
Social security – Turkish	6669	2025

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

### 3.2 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.

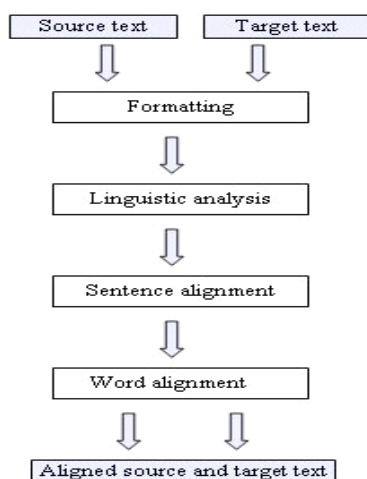


Figure 1: The modules of corpus annotation.

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.

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 Trigrams 'n' Tags part of speech tagger (Brants, 2000), trained on Swedish (Megyesi, 2002) 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. 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 and Türe, 2006). 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 account for 24.9% of all errors.

```

<s id="s11.4">
<w pos="DT_UTR_SIN_IND" head="3" deprel="DET" id="w11.4.1">Nâgon</w>
<w pos="JJ_POS_UTR_SIN_IND_NOM" head="3" deprel="DET" id="w11.4.2">annan</w>
<w pos="NN_UTR_SIN_IND_NOM" head="4" deprel="SUB" id="w11.4.3">titel</w>
<w pos="VB_PRT_SFO" head="0" deprel="ROOT" id="w11.4.4">fanns</w>
<w pos="AB" head="4" deprel="ADV" id="w11.4.5">inte</w>
<w pos="MAD" head="4" deprel="IP" id="w11.4.6">.</w>
</s>
<s id="s10.5">
<w pos="+Adj" head="3" deprel="MODIFIER" id="w10.5.1">Başka</w>
<w pos="+Num+Card^DB+Noun+Zero+A3sg+Pnon+Nom" head="6" deprel="SUBJECT"
id="w10.5.2">bir</w>
<w pos="+Noun+A3sg+Pnon+Nom" head="6" deprel="OBJECT" id="w10.5.3">başlık</w>
<w pos="+Adj^DB+Verb+Zero+Past+A3sg" head="0" deprel="ROOT" id="w10.5.4">yoktu</w>
<w pos="+Punc" head="6" deprel="PUNC" id="w10.5.5">.</w>
</s>

```

Figure 2: An example of morphological and syntactic annotation in XCES format.

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. Currently, we manually correct the syntactic annotation in each language. Figure 2 illustrates an example taken from Orhan Pamuk's book *The White Castle*, showing the morphological and syntactic annotation from the formatter and analyzers for

the sentence “Some other title did not exist.” in Swedish and Turkish in XCES format.

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).

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.

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.

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.

### **3.3 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 3 below illustrates the Uplug Connector interface.

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.

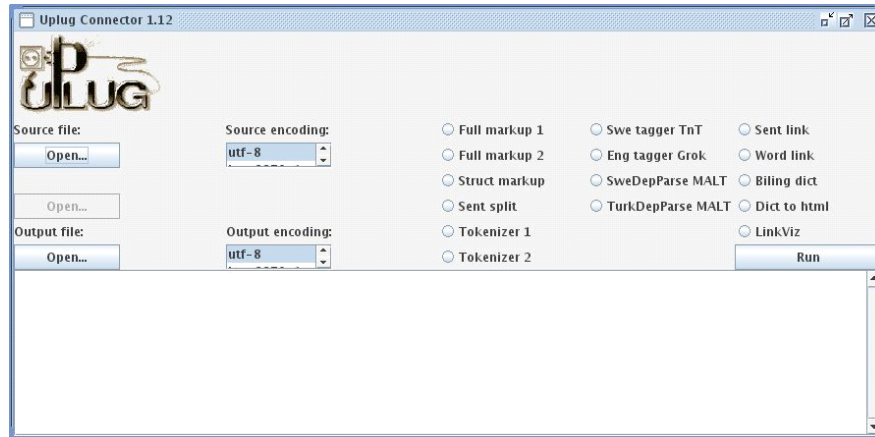


Figure 3: The Uplug Connector.

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 4 shows ISA with the aligned sentences taken from Orhan Pamuk's book *The White Castle*.

ISA & ICA / Interactive Sentence Alignment / vital1svtu << previous page | 10 | 20 | 50 | 100 | 200 | next page >>

>5 char <10 sentences cognates

XCF5 Align yourmail@host mail

change corpus link reset save align

s15.45	När jag sade att jag gjorde det, slapp jag äran och kunde även rädda en och annan av mina böcker.	Anladığını söyleyince hem küreğe verilmekten kurtuldum, hem de bu iki kitabımı kurtarmış oldum.	s19.6
s15.46	Men även det här privilegiet stod mig dyrt.	Ama bu ayrıcalığım da bana pahalıya pottadı.	s20.1
s15.47	Alla andra kaisina som sattes vid ärona föraktade mig.	Küreğe verilen öteki Hanıstayınlar hemen benden nefret etttiler.	s20.2
s15.48	Om de bara hade vågat skulle de ha dödat mig i lagerutrymmet som vi stängdes in i på nätterna, men de var rädda, eftersom jag genast hade etablerat kontakt med turkarna.	Ellerinden gelise geceleri birlikte kapalı olduğumuz ambarıda öldürülerdi beni, ama Türklerle hemen diğki kurdüğüm için korkuyorlardı da.	s20.3
s15.49	Vår fega kapten som spelets på på hade myse dött och som ett vanande exempel hade de slurit av näsan och öronen på dem som hade psikat slavarna och sankt ner dem i havet, på en liten flotte.	Kazığa oturtuldu korkak kapitanımız yeni ölünüşü, karlıcağları, bizimmi kulağım kesp ibret olsun diye bir sala koyup denize bırakmışlardı.	s20.4
s15.50	När jag behandlat sären på några turkiska soldater, mer med hjälp av mitt föränft än mina kamskapet i anatoni, och när de dessutom tillfrisknade av sig själva, trodde alla på att jag var läkare.	Anatoni bilgim değil de, aldimu kullanarak tedavi ettiğim birkaç Türk'ün yarasa kenellüğünden kapamınca herkes hekim olduğuma inanıdı.	s20.5
s15.51	Även några av mina avundsjuka fiender, som till turkarna sade att jag inte var läkare, visade mig sina sår i lagerutrymmet på nätterna.	Türlere hekim olmadığımı söyleyen bazı kaskacı düşmanlarımı bile geceleri ambarıda bana yaralarını gösterdiler.	s20.6
s15.52	Vi seglade in till Istanbul med pompa och ståt.	Istanbul'a gösterişli bir törenle girdik.	s21.1
s15.53	Sultanen, ännu ett barn, lär ha beskådat frändet.	Cocuk padişah bizi seyrediyomuş.	s21.2
s15.54	De lussade upp fanor i alla master, under dem hängde våra flaggor, ikoner med moeder Maria på och korsen upptoch de vida ande och de lit stadens hettevade män skjuta pålar på dem.	Bütün direklerin tepesine sancıklar çektiler, altlarına da bizim bayrakları, Meryem Ana tasvirlerini, haçları tersinden asıp kullandılar.	s21.3
s15.55	Kanonstälvorna dånade mot himlen och fick marken att skaka.	Derken toplar yeni göğü inlemeye başladı.	s21.4
s15.56	Frändet likom så många andra jag senare skulle få vara med om och mestadels beskåda från land med sorg och tristess men	Sonraları, birçokğum karadan bütüm, bildiğim ve neşeyle seyrettiğim tören çok uzun sürdü, güneşten baylanlar oldu.	s21.5

Figure 4: ISA showing the aligned sentences from *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 as shown in figure 5. 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.

SL6	»Att tänka sig att en person som förbryllar oss, har tilltråde till ett sätt att leva som är okänt och som känns mera attraktivt för dess mystik, att tro att vi kommer att börja leva endast genom dennes kärlek - vad annat är det, än början på en stor passion? «	" Alakamızı uyandıran bir kimseyi, bizce meçhul ve meçhullüğü derecesinde cazibeli bir hayatın unsurlarına karışmış sanmak ve hayata ancak onun sevgisiyle girebileceğimizi düşünmek bir aşk başlangıcından başka neyi ifade e"
-----	---	---

+Noun+A3sg+Pnon+Nom

Figure 5: 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 UplugConnector. This takes as input the text file with word link information produced by Uplug, see figure 6, and shows the word-pair frequencies. This visualization actually presents a bilingual lexicon created from the source and target language data.

Sofies värld

Nr	Frekvens	Svenska	Turkiska
1	62	"	"
2	58	.	.
3	58	?	?
4	34	,	,
5	29	och	ve
6	23	Sofie	Sofie
7	18	Men	Ama
8	17	en	bir
9	14	!	!
10	14	:	:

Figure 6: HTML-visualization of word alignment.

For the visualization and correction of the parallel syntactic trees, we choose Stockholm Tree Aligner (Lundborg, et al., 2007).<sup>1</sup> The tool allows the user to create links between corresponding nodes in two treebanks, hence allowing word and phrase alignment correction between our languages. The tool also contains a search function that implements the TigerSearch Query Language with additions for searching alignments. The visualization with Stockholm Tree Aligner for 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.

## 4 Further Developments

In the near future, we are going to apply the automatic annotation procedure on other languages of different types, such as Hindi and Persian and study the differences between the language pairs and the effects on the construction of parallel treebanks.

---

<sup>1</sup>See <http://www.ling.su.se/dali/downloads/trealigner/index.htm>.

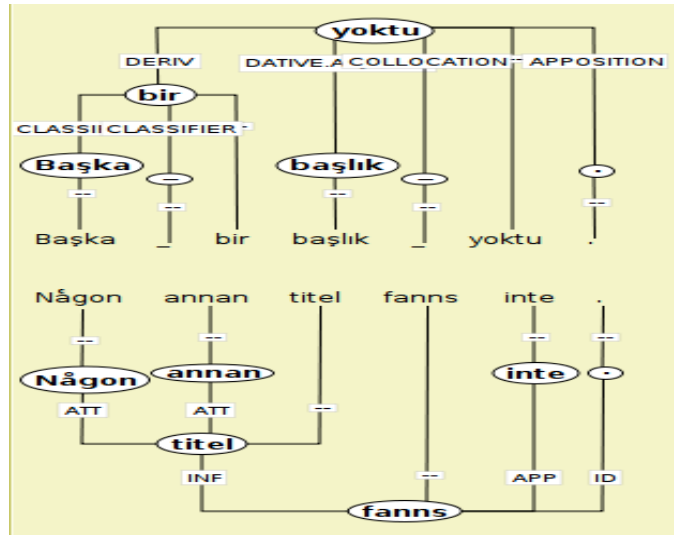


Figure 7: Dependency relations in Turkish and Swedish for the same sentence.

## 5 Conclusion

We have presented a Swedish-Turkish parallel treebank – a less processed language pair – containing approximately 165,000 tokens in Swedish, and 140,000 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 Eryiğit for the morphological annotation of Turkish. 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

Abeillé, A. (ed.) (2003). *Building and Using Parsed Corpora*. Text, Speech and Language Technology. Kluwer, Dordrecht.

- Ahrenberg, L. (2007). LinES: An English-Swedish Parallel Treebank. In *Proceedings of Nordiska Datalingvistdagarna*, NODALIDA 2007, Tartu, Estonia.
- Ahrenberg, L., M. Merkel, and M. Andersson (2002). A system for incremental and interactive word linking. In *Proceedings from The Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, 2002, pp. 485-490.
- Brants, T. (2000) TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*. Seattle, USA.
- Buchholz, S., and E. Marsi (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164.
- Church, K. W. (1993). Char align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, ACL.
- Cyrus, L., H. Feddes, and F. Schumacher (2003). FuSe – A Multi-Layered Parallel Treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, 14-15 November 2003, Växjö, Sweden.
- Erjavec, T. (2002). The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics*, 7(1), pp.1-20, 2002.
- Gale, W. A. and K. W. Church (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1), 75-102.
- Gustafson-Capková, S., Y. Samuelsson, and M. Volk (2007). *SMULTRON* (version 1.0) – The Stockholm MULtilingual parallel Treebank. An English-German-Swedish Parallel Treebank with Subsentential Alignments. <http://www.ling.su.se/dali/research/smultron/index.htm>.
- Hajic, J., E. Hajicová, P. Pajas, J. Panevová, P. Sgall, and B. Vidová-Hladká (2001). *Prague Dependency Treebank 1.0* (Final Production Label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0, 2001.
- Hpricot. A Fast, Enjoyable HTML and XML Parser for Ruby <http://code.whytheluckystiff.net/hpricot/> 2006.
- Ide, N. and G. Priest-Dorman. 2000. Corpus Encoding Standard – Document CES 1. Technical Report, Dept. of Computer Science, Vassar College, USA and Equipe Langue et Dialogue, France.

- Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Information Sciences Institute, University of Southern California.
- Lezius, W. (2002). TIGERSearch - Ein Suchwerkzeug für Baumbanken (German) in: Stephan Busemann (editor): Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002), Saarbrücken.
- Lundborg, J., T. Marek, M. Mettler, M. Volk (2007). Using the Stockholm TreeAligner. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*. Editors: Koenraad De Smedt, Jan Hajič and Sandra Kübler. NEALT Proceedings Series, Vol. 1 (2007), 73-78. © 2007 The editors and contributors. Published by Northern European Association for Language Technology (NEALT)
- MacIntyre, R. (1995). Penn Treebank tokenization on arbitrary raw text. <http://www.cis.upenn.edu/~treebank/tokenization.html>. University of Pennsylvania
- Megyesi, B. (2002). *Data-Driven Syntactic Analysis – Methods and Applications for Swedish*. PhD Thesis. Kungliga Tekniska Högskolan. Sweden.
- Megyesi, B. B., A. Sågvall Hein, and E. Csato Johanson (2006). Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.
- Megyesi, B. B., A. Sågvall Hein, and E. Csato Johanson (2006). Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.
- Megyesi, B. B. and B. Dahlqvist (2007). A Turkish-Swedish Parallel Corpus and Tools for its Creation. In *Proceeding of Nordiska Datalingvistdagarna, NODALIDA 2007*.
- Megyesi, B. B., B. Dahlqvist, E. Pettersson, and J. Nivre (2008). Swedish Turkish Parallel Treebank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Morocco.
- Nivre, J., J. Hall and J. 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)*, 2216–2219.
- Nivre, J., J. Hall and J. 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)*, 1392–1395.
- Oflazer, K. (1994). Two-level Description of Turkish Morphology. *Literary and Linguistic Computing*, **9**:2.
- Oflazer, K., B. Say, and Hakkani-Tür (2003). *Building a Turkish Treebank*. In Anne Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*, Kluwer, 261–277.
- Och, F. J. and H. Ney (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume **29**:1, pp. 19-51.
- Oksefjell, S. (1999). A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments. *International Journal of Corpus Linguistics*, **4**:2, 197-219.
- Resnik, P., M. Broman Olsen and M. Diab (1999). The Bible as a Parallel Corpus: Annotating the “Book of 2000 Tongues”. *Computers and the Humanities*, **33**:1-2, pp. 129-153, 1999.
- Samuelsson, Y. and M. Volk (2006). Phrase alignment in parallel treebanks. In Jan Hajič and Joakim Nivre, eds. *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories*, pp. 92-101, Prague.
- Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. 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)*. Genoa, Italy, 24-26 May 2006.
- SUC. Department of Linguistics, Umeå University and Stockholm University. 1997. SUC 1.0 Stockholm Umeå Corpus, Version 1.0. ISBN:91-7191-348-3.
- Tiedemann, J. (2003). *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Applications in Natural Language Processing*. PhD Thesis. Uppsala University.
- Tiedemann, J. (2004). Word to word alignment strategies. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland, August 23-27.
- Tiedemann, J. and L. Nygaard (2004). The OPUS corpus – parallel & free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal, May 26-28,

2004.

Tiedemann, J. (2005). Optimisation of Word Alignment Clues. In *Journal of Natural Language Engineering*, Special Issue on Parallel Texts, Rada Mihalcea and Michel Simard, Cambridge University Press.

Tiedemann, J. (2006). ISA & ICA – Two Web Interfaces for Interactive Alignment of Bitext. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Yuret, D. and F. Türe (2006). Learning morphological disambiguation rules for Turkish. In *Proceedings of HLT NAACL 2006*, pages 328-334, New York, NY.