The Swedish-Turkish Parallel Corpus and Tools for its Creation

Beata B. Megyesi & Bengt Dahlqvist
Department of Linguistics and Philology
Uppsala University
(beata.megyesi|bengt.dahlqvist)@lingfil.uu.se

Abstract

We present a Swedish-Turkish parallel corpus and the automatic annotation procedure with tools that we have been using in order to build the corpus efficiently. The method presented here can be transferred directly to build other parallel corpora.

1 Introduction

Parallel corpora containing texts and their translations have been a popular research area within natural language processing during the last decade. This is due to the fact that parallel corpora are very useful in language research allowing empirical studies, and various applications in natural language processing. In the past years, methods have been developed to build parallel corpora by automatic means, and to re-use translational data from such corpora for several applications, such as machine translation, multi-lingual lexicography, and cross-lingual domain-specific terminology.

In this paper, we describe a Swedish-Turkish parallel corpus, and the method and tools used for building it. Our primary goal is to build a representative language resource for Swedish and Turkish to be able to study the relations between these languages. The components of the language resource are texts that are in translational relation to each other and are analyzed linguistically. More specifically, our goal is to build a Swedish-Turkish parallel corpus with contrastive studies in focus.

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. We choose 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 corpus is part of the project “Supporting research environment for minor languages” aiming at building various types of language resources for Turkish, Hindi and Classic languages. The Swedish-Turkish corpus serves as a pilot project for building corpora 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 Swedish-Turkish parallel corpus is intended to be used in teaching, research, and applications such as machine translation.

The paper is organized as follows: Section 2 gives an overview of parallel corpora; Section 3 describes the corpus data while Section 4 presents the method for building the corpus and the tools used. In Section 5, we suggest some further improvements and lastly, in Section 6, we summarize the paper.

2 Parallel Corpora

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 above 20 languages and 8,000 docu-

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 136–143

ments of legislative text, covering a variety of domains. 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 IJS-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 for languages other than European or that exclude English are rare. There is therefore a need to develop language resources in general, and parallel corpora in particular for other language pairs as well.

Next, we describe the development of a Swedish-Turkish parallel corpus. To our knowledge, there is no similar or comparable resource such as the corpus we present in this paper.

3 Corpus Content

The corpus consists of original texts and their translations from Turkish to Swedish and from Swedish to Turkish with the exception of one text which is a translation to both languages.

We collected written texts to build a balanced corpus with respect to translational direction. The corpus contains both fiction and technical documents. The fiction part consists of one full novel “The White Castle” by Orhan Pamuk, and the first chapter of “Sofie’s world” by Jostein Gaardner. As for the non-fiction, a book “Islam and Europe” by Ingmar Karlsson, a booklet “Information from the Swedish Migration office” and a number of short information brochures for Turkish immigrants from Swedish governmental agencies are included.

In Table 1, the corpus material is summarized. In total, the corpus consists of approximately 150,000 tokens in Swedish and 126,000 tokens in Turkish. Divided into text types, the fiction part of the corpus includes 59,720 tokens in Swedish, and 41,484 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 1. The corpus data divided into text categories with number of tokens and types.

Document	***# Token***	***# Type***
Fiction		
The White Castle - Swe	53232	7748
The White Castle - Tur	36684	12472
Sofie's world - Swe	6488	1466
Sofie's world - Tur	4800	2215
Non-fiction		
Islam and Europe - Swe	55945	10977
Islam and Europa - Tur	48893	14128
Info about Sweden - Swe	24107	4576
Info about Sweden - Tur	23660	7119
Retirement - Swe	3417	818
Retirement - Tur	3664	1188
Dublin - Swe	392	169
Dublin - Tur	394	230
Pregnancy - Swe	949	409
Pregnancy - Tur	1042	567
Psychology - Swe	347	193
Psychology - Tur	281	220
Movement - Swe	543	300
Movement - Tur	568	369
Social security - Swe	5201	846
Social security - Tur	6669	2025

4 Corpus Annotation Procedure

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.

The following steps below give an overview of the annotation procedure and the involved tools.

1. **Preprocessing** for cleaning up the original files, partly manually.
2. **UplugConnector** for markup, linguistic analysis and alignment in a graphical interface to the **Uplug toolkit** (Tiedemann, 2003).
3. **ISA** (Tiedemann, 2006) for visualization of sentence alignment and manual correction.
4. Visualization of the material with the linguistic analysis without showing the structural markup using **Hpricot**.
5. **ICA** (Tiedemann, 2006) for visualization of the word alignment.

4.1 Preprocessing

First, the original materials 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 scanned and proof-read the material and, where necessary, corrected it to ensure that the plain text file is complete and correct.

Then, the texts are encoded according to international standards by using UTF-8 (Unicode). The plain text files are then processed by various tools. The sentences of the formatted texts in the source and target language are linguistically analyzed and aligned automatically, and the words are linked to each other in the two languages. Next, the corpus architecture and tools used to build the corpus is presented in more detail.

4.2 Corpus Markup

The clean plain text files are processed to markup the data, to annotate it with morpho-syntactic features, and to align the texts on the paragraph, sentence and word level. For this purpose, we use the Uplug toolkit which is a collection of tools for processing corpus data, created by Jörg Tiedemann (2003). Uplug was developed for word alignment in parallel corpora and utilizes BLARKs where possible. Uplug can be used for sentence splitting, tokenization, tagging by using external taggers, and paragraph, sentence and word alignment. Figure 1 gives an overview of the main modules in the corpus annotation procedure with Uplug.

All tools included are freely available for research purposes and are built in as components in the Uplug toolkit.

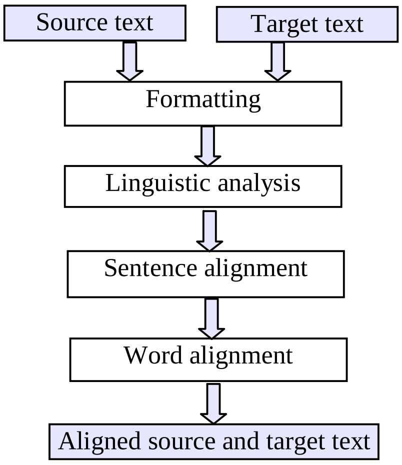
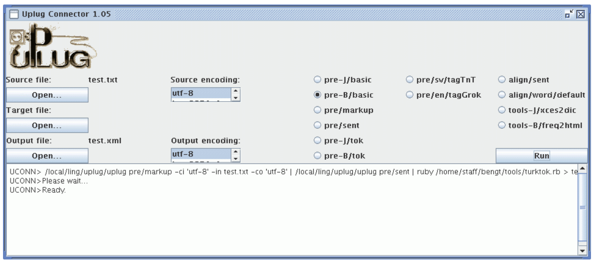


Figure 1. Modules of Uplug

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. The figure below illustrates the Uplug Connector interface.



Source file: test.txt
Source encoding: utf-8
Target file:
Output file: test.xml
Output encoding: utf-8

○ pre-J/basic
● pre-B/basic
○ pre/markup
○ pre/sent
○ pre-J/tok
○ pre-B/tok
○ pre/sv/tagTnT
○ pre/en/tagGrok
○ align/sent
○ align/word/default
○ tools-J/xces2dic
○ tools-B/freq2html

```
UCONN> /local/ling/uplug/uplug pre/markup -ci 'utf-8' -in test.txt -co 'utf-8' | /local/ling/uplug/uplug pre/sent | ruby /home/staff/bengt/tools/turktok.rb > te
UCONN>Please wait...
UCONN>Ready.
```

Figure 2. The Uplug Connector

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. In the toolkit, the user can also call for the sentence and word aligners and their visualization tool.

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 to another resource if the available ones do not fit his/her needs, for example an external tokenizer, sentence splitter, or tagger.

4.2.1 Formatting

Each part of the corpus is clearly marked and annotated. We use the international 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 sentence splitter is used to break the texts into sentences, and the texts are tokenized for both languages. Since the default tokenizer in Uplug (to our knowledge) does not handle character entities and hyphens in Turkish words correctly, an alternative tokenizer was developed in the project, loosely based on the Penn Treebank tokenizer by Robert MacIntyre (1995).

The sentences and words are then marked as s and w respectively, and receive an identification number. An example taken from Orhan Pamuk's book "The White Castle" is shown below for the sentence "Some other title did not exist" first in Swedish "Någon annan titel fanns inte.",

```
<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>
<w id="w11.4.6">.</w>
</s>
```

then in Turkish "Başka bir başlık yoktu." :

```
<s id="s10.5">
<w id="w10.5.1">Başka</w>
<w id="w10.5.2">bir</w>
<w id="w10.5.3">başlık</w>
<w id="w10.5.4">yoktu</w>
<w id="w10.5.5">.</w>
</s>
```

4.2.2 Linguistic analysis

Once the sentences and tokens are identified, the data is analyzed linguistically. For the linguistic annotation, external morphological analyzers and part-of-speech taggers are used for the specific languages.

The Swedish texts are annotated with the Trigrams'n'Tags PoS tagger (Brants, 2000). The tagger was trained on Swedish (Megyesi, 2002) using the Stockholm-Umeå Corpus (SUC, 1997). For the labels, we use the PAROLE annotation scheme developed for Swedish (Ejerhed and Ridings, 1995). The tokens are annotated with part-of-speech and morphological features and are disambiguated according to the syntactic context with an accuracy of approximately 96% (Megyesi, 2002). An example of the morphological annotation for the same sentence as previously is shown below.

```
<s id="s11.4">
<w pos="DI@US@S" id="w11.4.1">Någon</w>
<w pos="AQPUSNIS" id="w11.4.2">annan</w>
<w pos="NCUSN@IS" id="w11.4.3">titel</w>
<w pos="V@IISS" id="w11.4.4">fanns</w>
<w pos="RG0S" id="w11.4.5">inte</w>
<w pos="FE" id="w11.4.6">.</w>
</s>
```

The Turkish material is analyzed linguistically 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 morphological analyzer does not disambiguate the tokens. Preliminary results show on part of the Turkish material that 74% of the tokens were correctly and completely analyzed with morphological features. The rest of the tokens are either ambiguous, or are unknown, often foreign words.

4.2.3 Sentence alignment, visualization and correction

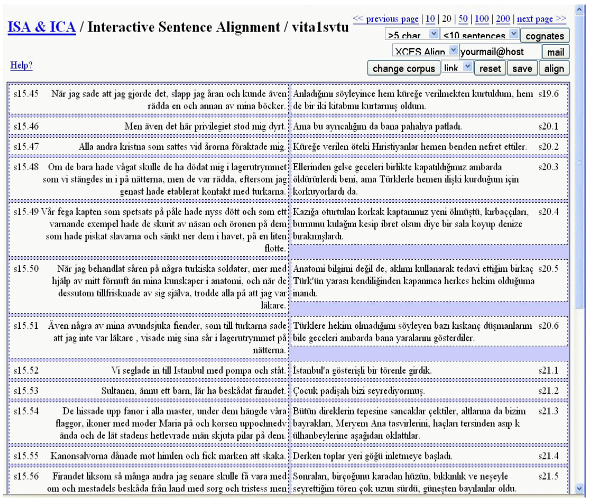
Aligning the translated segments with source segments are essential for building parallel corpora. We use standard techniques for the establishment of links between source and target language segments. Paragraphs and sentences are aligned by using the length-based approach developed by Gale and Church (1993).

The aligned sentences are stored in XML format, as shown in the example below.

```
<cesAlign toDoc="vt.xml" version="1.0" fromDoc="vs_tnt.xml">
 <linkGrp targType="s" toDoc="vt.xml" fromDoc="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"/>
...
```

As the XML representation of the linking result is not user friendly even for people used to this kind of annotation, an interface for the visualization of the alignment result is required. In addition, since the automatic alignment generates some errors, we also need an interface 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”.


ISA & ICA / Interactive Sentence Alignment / vita1svtu
<< previous page | 10 | 20 | 50 | 100 | 200 | next page >>
>5 char
<10 sentences
cognates
XCES Align
yourmail@host
mail
Help?
change corpus
link
reset
save
align

s15.45	När jag sade att jag gjorde det, slapp jag åran och kunde även rädda en och annan av mina böcker.	Anladığımı söyleyince hem küreğe verilmekten kurtuldum, hem de bir iki kitabımı kurtarmış oldum.	s19.6
s15.46	Men även det här privilegiet stod mig dyrt.	Ama bu ayrıcalığım da bana pahalıya patladı.	s20.1
s15.47	Alla andra kristna som sattes vid årorna föraktade mig.	Küreğe verilen öteki Hıristiyanlar hemen benden nefret ettiler.	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 gelse geceleri birlikte kapatıldığımız ambarda öldürürlerdi beni, ama Türklerle hemen ilişki kurduğum için korkuyorlardı da.	s20.3
s15.49	Vår fega kapten som spetsats på påle hade nyss dött och som ett varnande exempel hade de skurit av näsan och öronen på dem som hade piskat slavarna och sänkt ner dem i havet, på en liten flotte.	Kazığa oturtulan korkak kaptanımız yeni ölmüştü, kırbaççıları, burnunu kulağını kesip ibret olsun diye bir sala koyup denize bırakmışlardı.	s20.4
s15.50	När jag behandlat såren på några turkiska soldater, mer med hjälp av mitt förnuft än mina kunskaper i anatomi, och när de dessutom tillfrisknade av sig själva, trodde alla på att jag var läkare.	Anatomi bilgimi değil de, aklımı kullanarak tedavi ettiğim birkaç Türk'ün yarası kendiliğinden kapanınca herkes hekim olduğuma inandı.	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.	Türklere hekim olmadığımı söyleyen bazı kıskanç düşmanlarım bile geceleri ambarda bana yaralarını gösterdiler.	s20.6
s15.52	Vi seglade in till Istanbul med pompa och ståt.	İstanbul'a gösterişli bir törenle girdik.	s21.1
s15.53	Sultanen, ännu ett barn, lär ha beskådat firandet.	Çocuk padişah bizi seyrediyormuş.	s21.2
s15.54	De hissade upp fanor i alla master, under dem hängde våra flaggor, ikoner med moder Maria på och korsen uppochnedv ända och de lät stadens hetlevrade män skjuta pilar på dem.	Bütün direklerin tepesine sancaklar çektiler, altlarına da bizim bayrakları, Meryem Ana tasvirlerini, haçları tersinden asıp k ülhanbeylerine aşağıdan oklattılar.	s21.3
s15.55	Kanonsalvorna dånade mot himlen och fick marken att skaka.	Derken toplar yeri göğü inletmeye başladı.	s21.4
s15.56	Firandet liksom så många andra jag senare skulle få vara med om och mestadels beskåda från land med sorg och tristess men	Sonraları, birçoğunu karadan hüzün, bıkkınlık ve neşeyle seyrettiğim tören çok uzun sürdü, güneşten bayılanlar oldu.	s21.5

Figure 3. ISA showing the aligned sentences from “The White Castle”.

Once the sentences are aligned in the source and target language, we send it for manual correction to a student who speaks both languages. With the help of ISA, the manual correction is easy and fast.

The results we present below are based on the sentence alignment results for the first chapter of the novel “The White Castle” by Orhan Pamuk.

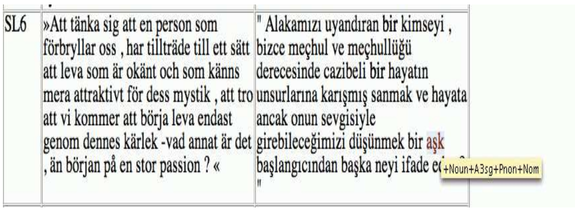
The manually corrected alignment resulted in 178 sentence pairs after merges and splits. The distribution of the alignment types after the manual alignment is shown in column two in Table 2.

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

Link type: Swedish-Turkish	***Manual Number***	***Automatic***	
		Number	***Correct (%)***
1-0	9	0	0
1-1	144	126	110 (87.3)
1-2	3	3	3 (100)
2-1	15	39	12 (33.0)
3-1	7	0	0
Total	**178**	**168**	**125 (74.4)**

The Uplug automatic sentence alignment produced 168 sentence pairs. The correctness of these compared with the manual alignment is presented in column three and four in Table 2. 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.

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 is displaying the sentences aligned together with the linguistic information for each word shown in pop-up windows.



SL6	»Att tänka sig att en person som förbryllar oss , har tillträde 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 ed "

Figure 4. Visualization of aligned sentence pairs with linguistic annotation shown in the pop-up window.

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

4.2.4 Word alignment and visualization

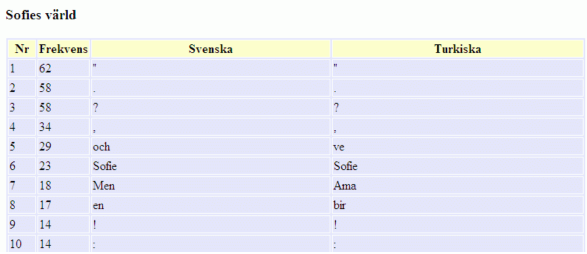
As the next step, words and phrases are aligned using the clue alignment approach (Tiedemann, 2003), and the toolbox for statistical machine translation GIZA++ (Och and Ney, 2003), also implemented in Uplug.

Results show that the word aligner aligned approximately 69% of the words correctly. For a pilot 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 (padişaha). 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 with 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 sentence alignment.

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 5, and shows the word-pair frequencies. This visualization in fact serves as 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 5. HTML-visualization of word alignment.

5 Further Developments

In the near future, we would like to extend the linguistic analysis with syntactic features for both languages, and apply a better morphological analysis for Turkish sentences. Also, we plan to use these annotations to improve the automatic word alignment, and use an appropriate tool for visualizing the syntactic annotation. In this way, we easily can build a parallel treebank. Finally, manual corrections of all materials in the corpus are carried out.

6 Conclusions

We presented a Swedish-Turkish parallel corpus − a less processed language pair − containing approximately 150,000 tokens in Swedish, and 126,000 tokens in Turkish. The corpus 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 annotation of the involved languages. The corpus 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 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

Thorsten Brants. 2000. TnT − A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*. Seattle, USA.

Kenneth W. Church. 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.

Eva Ejerhed and Daniel Ridings. 1995. *Parole ->SUC and SUC -> Parole*. http://sprakdata.gu.se/lb/sgml2suc.html

Tomaž Erjavec. 2002. The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics, 7(1),* pp.1-20, 2002.

William A. Gale, and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics, 19(1),* 75-102.

Hpricot. A Fast, Enjoyable HTML and XML Parser for Ruby http://code.whytheluckystiff.net/hpricot/ 2006.

Nancy Ide, and Greg Priest-Dorman. 2000. *Corpus Encoding Standard – Document CES 1*. Technical Report, Dept. of Computer Science, Vassar College, USA and Equipe Langue et Dialogue, France.

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

Robert MacIntyre. 1995. *Penn Treebank tokenization on arbitrary raw text*. University of Pennsylvania, http://www.cis.upenn.edu/~treebank/tokenization.html

Beata Megyesi. 2002. *Data-Driven Syntactic Analysis − Methods and Applications for Swedish.* PhD Thesis. Kungliga Tekniska Högskolan. Sweden.

Beata B. Megyesi, Anna Sågvall Hein, and Eva Csato Johanson. 2006. Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy.

Kemal Oflazer. 1994. Two-level Description of Turkish Morphology, Literary and Linguistic Computing, Vol. 9, No:2.

Franz Josef Och, and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume 29:1, pp. 19-51, March 2003.

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 Humanities, 33(1-2),* pp. 129-153, 1999.

John Sinclair. (Ed.) 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel 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.

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

Jörg Tiedemann. 2004. Word to word alignment strategies. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004). Geneva, Switzerland, August 23-27.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus − parallel & free. In *Proceedings of the Fourth International Conference on Language Resources* *and Evaluation (LREC'04)*. Lisbon, Portugal, May 26-28, 2004.

Jörg Tiedemann. 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.

Jörg Tiedemann. 2006. ISA & ICA – Two Web Interfaces for Interactive Alignment of Bitext. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy.