Building a Swedish-Turkish Parallel Corpus

Beáta Bandmann Megyesi, Anna Sågvall Hein, Éva Csató Johanson

Department of Linguistics and Philology, Uppsala University
Box 635, SE-751 26 Uppsala, Sweden
[beata.megyesi | anna | eva.csato-johanson]@lingfil.uu.se

Abstract

We present a Swedish-Turkish Parallel Corpus aimed to be used in linguistic research, teaching, and applications in natural language processing, primarily machine translation. The corpus being under development is built by using a Basic LAnguage Resource Kit (BLARK) for the two languages which is then used in the automatic alignment phase to improve alignment accuracy. The corpus is balanced with respect to source and target language and is automatically processed using the Uplug toolkit.

1. Introduction

Language resources such as carefully compiled corpora of collected texts and utterances play a central roll in language studies and natural language processing. Pioneering projects such as the COBUILD project (Sinclair, 1987) have demonstrated the significance of authentic language material collected in large corpora for capturing information about the language. Moreover, parallel corpora including texts and their translations contain highly valuable linguistic data across languages. Recently, methods and technologies have been developed for re-using translational data from such corpora for multi-lingual lexicography, cross-lingual and domain-specific terminology, computer-aided translation and machine translation.

The aim of the project presented in this paper is to build a representative language resource for Swedish and Turkish in order to be able to study the relations between these languages. The components of the language resource will be texts that are in translational relation to each other as well as tools for the automatic analysis of these languages.

More specifically, the goal is to build and annotate a Swedish-Turkish Parallel Corpus automatically by using a basic language resource kit (BLARK) for the particular languages which can then be used in the automatic alignment phase to improve alignment accuracy. The parallel corpus is intended to be used in linguistic research, teaching and applications such as machine translation.

2. Parallel Corpora

A parallel corpus, sometimes called bitext, is a collection of original texts translated to another language where the texts, paragraphs, and sentences down to word level are typically linked to each other.

Parallel corpora are of great importance in language studies, teaching and many natural language processing applications such as machine translation, cross language information retrieval, word sense disambiguation, bilingual terminology extraction as well as induction of tools across languages.

One of the most 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 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 Sloven-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 not very common. There is therefore a need to develop language resources, such as parallel corpora 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. The Swedish-Turkish Corpus

Before we present the corpus data, we give a short overview of the involved languages as they are less known, and belong to different language types.

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 Data

It is now a well recognized fact that a corpus is more than just a collection of electronic texts. Corpus data have to be selected with care with respect to the intended applications. In this project we emphasize quality with regard to content and translation. We focus on a collection of written texts to build a balanced corpus of the source and target language. As for genre, we choose both fiction and technical documents.

The corpus consists, so far, of the texts listed below.

Fiction

- Orhan Pamuk: Beyaz Kale
  Swedish title: Den vita borgen
  English title: The White Castle
  Source language: Turkish (36,626 words),
  Target language: Swedish (53,241 words)

- Jostein Gaardner: Sofies verden, chapter 1 and 2
  Swedish title: Sofies värld
  Turkish title: Sofie'nin dünyasi
  English title: Sofie's world
  Source language: Norwegian
  Target language: Swedish (7,926 words) and Turkish (5,307 words)

Special/Technical texts

- Ingmar Karlsson: Islam och Europa
  Turkish title: Islam Ve Avrupa
  English title: Islam and Europe
  Source language: Swedish (56,232 words)
  Target language: Turkish (49,258 words)

- Sverige Information (Information from the Swedish Migration Office)
  Source language: Swedish (23,859 words)
  Target language: Turkish (23,562 words)

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 near future.

4. Corpus Annotation Procedure

The corpus material is processed automatically, partly by using the Uplug toolkit which is a collection of tools for processing corpus data, developed 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 shows the main modules of the corpus annotation procedure.

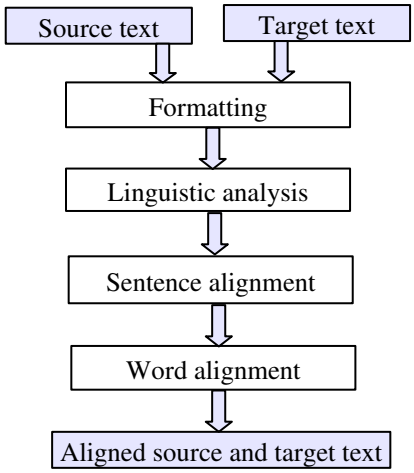

Figure 1. Modules in the corpus annotation procedure

We start the annotation by cleaning up the original material that we received from the different publishers. This means that the various formats, for example rtf, doc, and pdf, 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.

As the next step, the texts are encoded according to international standards by using UTF-8 (Unicode) and ISO-8859-1 for Latin-1 (which includes Swedish) and ISO-8859-9 for Latin-5 (which includes Turkish). The reason for keeping the text in various encodings is that some linguistic analyzers cannot handle Unicode yet.

The plain text files are then 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. Thereafter the tokens are annotated with their part-of-speech including morphological features by using two different taggers for the two languages. Parts of the Swedish material has also been parsed by a rule-based phrase structure parser SPARKparse (Megyesi, 2002).

The paragraphs and sentences of the formatted texts in the source and target language are aligned

automatically and the words are linked in the two languages by the Uplug kit. Next, the corpus architecture is presented in more detail.

4.1. Annotation Format

Properties of each part of the corpus is clearly marked and annotated. For the annotation format we use international XML Corpus Encoding Standard (XCES). The sentences and words are clearly marked and receive an id-number, as shown below for the sentence "Islam is not a new phenomenon in European territory." first in Swedish, then in Turkish.

```
<s id="s7.29">
 <w id="w7.29.1">Islam</w>
 <w id="w7.29.2">är</w>
 <w id="w7.29.3">ingen</w>
 <w id="w7.29.4">ny</w>
 <w id="w7.29.5">företeelse</w>
 <w id="w7.29.6">på</w>
 <w id="w7.29.7">europeisk</w>
 <w id="w7.29.8">mark</w>
 <w id="w7.29.9">.</w>
</s>

<s id="s24.31">
 <w id="w24.31.1">Ýslam</w>
 <w id="w24.31.2">,</w>
 <w id="w24.31.3">Avrupa</w>
 <w id="w24.31.4">topraklarýnda</w>
 <w id="w24.31.5">yeni</w>
 <w id="w24.31.6">bir</w>
 <w id="w24.31.7">olgu</w>
 <w id="w24.31.8">deðildir</w>
 <w id="w24.31.9">.</w>
</s>
```

4.2. Linguistic Analysis

The texts in both languages are analyzed with respect to their morphological structure and part-of-speech (PoS).

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) which is a balanced corpus, consisting of various genres (similar to the Brown corpus). For the labels, we use the PAROLE annotation scheme developed for Swedish (Ejerhed and Ridings, 1995). It consists of 163 different tags containing information about PoS and inflectional properties of the words. 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).

The Turkish material is analyzed with 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. We hope to be able to disambiguate the alternative analyzes soon by parsing the texts. 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 is either ambiguous, or is unknown, often foreign words.

4.3. Automatic Alignment

Essential for building parallel corpora is the alignment of translated segments with source segments. We use standard techniques for the establishment of links between source and target language segments. First, paragraphs and sentences are aligned by using the length-based approach by Gale and Church (1993). Next, words and phrases are aligned using the clue alignment approach (Tiedemann, 2003), and the toolbox for statistical machine translation GIZA++ (Och and Ney, 2003). These tools are freely available for research purposes and are built in as components in the Uplug toolkit (Tiedemann, 2003) described previously.

Preliminary results show that approximately 31% of the words were not correctly aligned at the word level. For a pilot evaluation of the results, we investigated the error level on 7077 word pairs in Swedish and Turkish (sorted by decreasing frequency) taken from "The White Castle" written by Orhan Pamuk.

Of the wrongly 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. Not surprisingly, many errors depend on the fact that Swedish has two or more words for expressions that constitute only one word in Turkish. 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 what 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. Other examples of alignment errors are due to erroneous formatting. One fairly commonly occurring error is when the aligner does not attach the apostrophe (') with the following suffix (attached to proper nouns in Turkish) to the Swedish preposition due to tokenization problem.

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.

Note that these results are preliminary and gives us only an indication of the error level and the types described above. We have to evaluate the sentence– and word alignment and study the cause of the error types in the rest of the material as well to improve alignment accuracy.

5. Further Development

Since the quality and the type of the linguistic analysis achieved for the two types of languages differs to a great extent, we aim to investigate how the more detailed and accurate features in one language (in our case Swedish) can help and improve the automatic alignment process and the linguistic annotation of the other language.

After aligning the material, an important issue is the manual correction of automatically aligned segments. For this we intend to develop appropriate tools that support the correction of erroneous links in the future.

Lastly, we intend to continuously include more translated materials to the corpus and, when possible, make it available to the public.

We believe that the method used for the development of the Swedish-Turkish parallel corpus can be fairly easily applied to other language pairs as well when creating parallel corpora of them.

6. Conclusions

In this paper, we presented a Swedish-Turkish parallel corpus of a new language pair where the languages belong to different language types. The corpus which consists of approximately 150,000 words in Swedish and 100,000 words in Turkish is balanced with respect to source and target language. The texts are annotated with part-of-speech and morphological features and are automatically aligned at sentence- and word-level.

The corpus is still under development, and we hope that we may enlarge it further with material that can be made freely available to the public.

Acknowledgments

We are very grateful to Jörg Tiedemann for providing us with the Uplug toolkit and assisting in the linking process when it was necessary. Also, a big thank to Kemal Oflazer and his staff at Sabanci University for providing us with the Turkish morphological analysis. Lastly, this project would not be possible without being financially supported by the Swedish Research Council and the Faculty of Languages at Uppsala University.

7. References

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

Church, K. (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.

Ejerhed, E. and Ridings, D. (1995) *Parole -> SUC and SUC -> Parole*. http://sprakdata.gu.se/lb/sgml2suc.html

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

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

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

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

Megyesi, B. (2002) *Data-Driven Syntactic Analysis - Methods and Applications for Swedish.* PhD Thesis. Kungliga Tekniska Högskolan. Sweden.

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

Och, F. J., and Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

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, Ph., Broman Olsen, M., and Diab, M. (1999) The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues". *Computers and the Humanities, 33(1-2)*, pp. 129-153, 1999.

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

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. (1999) Uplug - A Modular Corpus Tool for Parallel Corpora. In Lars Borin (ed.), 2002, *Parallel corpora, parallel worlds.* Selected papers from a symposium on parallel and comparable corpora at Uppsala University. Amsterdam: Rodopi.

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 Nygaard, L. (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.

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