USING PARALLEL CORPORA IN DATA-DRIVEN TEACHING OF TURKISH IN SWEDEN

Éva Á. Csató, eva.csato@lingfil.uu.se
Songül Kilimci, songul.kilimci@lingfil.uu.se
Beata Megyesi, beata.megyesi@lingfil.uu.se

Department of Linguistics and Philology, Uppsala University

Abstract

The paper demonstrates how data-driven learning methods are applied in teaching Turkish as a foreign language at the Department of Linguistics and Philology, Uppsala University. In data-driven teaching, language corpora, concordance programs, and annotation tools developed in collaboration with computational linguists are employed. This paper illustrates how resources developed initially for research purposes in different subjects (such as Computational Linguistics, Linguistics, Turkic languages), are now being used in teaching environments.

We present the Swedish-Turkish parallel corpus providing students and researchers with easily accessible annotated linguistic data. The web-based corpora can be used both by regular and distance students. They function also as learning tools for formulating and testing hypotheses concerning lexical, morphological and syntactic aspects of Turkish. Furthermore, they help the students to practice contrastive studies and translation between Swedish and Turkish.

Key Words: Parallel Corpora, Data-Driven Learning Method, Language Teaching, Computational Linguistics.

Introduction

Recently, different types of linguistic corpora have become a primary source in the study of languages. The word corpus is used to describe "a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study" (Hunston, 2002: 2). Corpora are stored and accessed electronically "a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written text or recordings, which have been collected for linguistic study" (Hunston, 2002: 2). Corpora are stored and accessed electronically and are often used "to give information about how a language works, to make comparisons between languages, to compare the use of apparent translations, to present the structure of a language" (Hunston, 2002: 13-14). Hunston gives several examples of the use of corpora to study, e.g. frequency, phraseology, usages of words or phrases, collocations, statistical tendency of words to co-occur (2002: 13-14).

Parallel corpora contain texts that have been translated from one language into another. In parallel texts, learners can find potential equivalent expressions in the given languages and investigate semantic, syntactic and collocational differences. Such investigations can lead to comparative language studies (Johansson 2007).

Data-driven learning (DDL) makes use of corpora, which allow learners to act as ‘language detectives’ (Johns, 1997:101). Students are inspired to discover themselves how words, idioms, phrases and grammatical phenomena are used in the target language. They can access the database and utilize information from the corpora (Dush, 2003). The authenticity of the data ensures that learners are presented with samples, which reflect the way people actually speak or write. Teachers can present regularities in the parallel texts to make learners aware of grammatical, lexical or discourse features, which distinguish their interlanguage from the target norm (Granger, 1998).

Less Commonly Taught Languages at Uppsala University

Over forty different languages are taught at the Faculty of Languages at Uppsala University. Of these, more than twenty, so-called less commonly taught languages, such as Arabic, Hindi, Kurdish, Farsi, Swahili, Turkish, and Uzbek, are taught in the Department of Linguistics and Philology. As crucial methodological issues are specific to the less commonly taught languages, American universities have created the National Council of Organizations of Less Commonly Taught Languages. The aim of this council is to develop a uniform set of goals, common guidelines for curricula, pooling teaching materials, and arranging conferences for exchange of experience. In Sweden, several of the less commonly taught languages are taught only at Uppsala University. Thus, there is a need in our department to pool our resources in order to meet the demands of teaching these languages.

In our language courses, students are expected to acquire not only practical skills in speech and writing. They are expected to study the target language from a linguistic point of view, i.e. they have to learn about the typological, grammatical, lexical properties of the language and acquire skills and methods in order to be able to carry out research on the language. We give some examples below to demonstrate how this multifaceted task can be achieved using the data-driven learning method. This method uses resources such as language corpora, concordance programs, and annotation tools – the tools developed in collaboration with the computational linguists. Resources developed initially for research purposes in different subjects (such as Computational Linguistics, Linguistics, Turkic languages), are now being used also in various teaching environments. Internet resources create a new social context for learning and enhance the students’ engagement (Kilimci, 2010). Here we will demonstrate one example - the use of parallel corpora.

The Turkish-Swedish Parallel Corpus

The Swedish-Turkish parallel corpus consists of original texts, both fiction and non-fiction and their translations provided by professional translators. In total, the corpus consists of 288,701 tokens in Swedish, and 162,302 tokens in Turkish. Table 1 below gives an overview of the corpus data with the number of tokens in the two languages.

<table>
<thead>
<tr>
<th>Type of Text</th>
<th>Swedish</th>
<th>Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td>The White Castle (O. Pamuk)</td>
<td>58,684</td>
<td>44,176</td>
</tr>
<tr>
<td>Sofie’s world (J. Gaardner)</td>
<td>7,393</td>
<td>5,651</td>
</tr>
<tr>
<td>The royal physician’s visit (PO EnQUIST)</td>
<td>20,780</td>
<td>16,983</td>
</tr>
<tr>
<td>Islam and Europe (I. Karlsson)</td>
<td>61,529</td>
<td>58,353</td>
</tr>
<tr>
<td>Info about Sweden (Migration Office)</td>
<td>26,649</td>
<td>28,139</td>
</tr>
<tr>
<td>Pregnancy and Giving Birth</td>
<td>1,076</td>
<td>1,221</td>
</tr>
<tr>
<td>Exercise and Food</td>
<td>616</td>
<td>685</td>
</tr>
<tr>
<td>Psychological Issues</td>
<td>355</td>
<td>339</td>
</tr>
<tr>
<td>Retirement</td>
<td>3,770</td>
<td>4,267</td>
</tr>
<tr>
<td>Dublin</td>
<td>451</td>
<td>469</td>
</tr>
<tr>
<td>UN Declaration of Human Rights</td>
<td>1,831</td>
<td>1,604</td>
</tr>
<tr>
<td>What is unicode</td>
<td>539</td>
<td>424</td>
</tr>
</tbody>
</table>
The corpus is automatically created by using and adjusting existing tools for the linguistic analysis, the automatic alignment, and visualization. Figure 1 shows the annotation procedure. The texts are first cleaned up and formatted into text files. Then, the words, punctuation marks and the sentences are segmented by a tokenizer and a sentence segmenter, developed separately for each language. The tokens are morphologically analyzed with part of speech and inflectional features. For the morphosyntactic annotation, external morphological analyzers and part-of-speech taggers are used for the two languages. The Swedish texts are annotated with the HunPoS tagger (Halácsy et al., 2007), an open source reimplementation of the Trigrams’n’Tags (TnT) tagger (Brants, 2000), with an average accuracy of 96% (Megyesi, 2008). The Turkish material is morphologically analyzed using a Turkish analyzer (Oflazer, 1994) and a disambiguator, which automatically learns morphological disambiguation rules from a decision list induction algorithm achieving an accuracy of approximately 96% (Yuret and Türe, 2006). The sentences are also syntactically annotated in both languages with dependency structures by using MaltParser (Nivre et al., 2006) trained on each language separately to build a syntactically annotated corpus, so called treebank. The sentences, phrases and words in the two languages are also aligned text by text. The output is linguistically analyzed parallel texts. The corpus is under development and partly manually corrected.

Figure 1. Annotation procedure

Learning Tools in the Turkish-Swedish Parallel Corpus

The aim of the Swedish-Turkish parallel corpus is to provide Swedish speaking students and researchers with easily accessible annotated linguistic data on Turkish. The web-based corpora can be used both by regular and distance students in their acquisition of new vocabulary items and their usage. They function also as a learning platform for testing hypotheses concerning the morphological and syntactic aspects of Turkish grammar. Further, they help the students to practice translation between Swedish and Turkish. All that is possible due to the fact that the Swedish-Turkish parallel corpus is available in annotated form. The annotations, on request, are visualized in pop-up windows, as shown in Figure 2.

Figure 2. Morphological analysis in a pop-up window

The interface for displaying syntactic information is not ready yet. A search tool assists the students to create concordance lists. They can search for whole words, beginnings of words, and parts of words or ends of words in Turkish or Swedish. The concordance lists display whole sentences in which the target item appears and it is highlighted. The selected sentences are aligned with their translational equivalents, as illustrated in Figure 3.
Figure 3. Concordance list: Search results for the word ‘insan’

This form of displaying the linguistic data is much more suitable for learning than KeyWords In Context (KWIC) lists in which only the immediate environment of the target item is shown. Such lists are used to find frequent patterns of usage, transformational equivalents, different meanings of polysemic words, translational equivalents of Turkish grammatical categories, etc. Different types of exercises are designed and published on the Internet.

In a data-driven learning approach, students use corpora directly in their own learning. They use the corpora, for example, to discover linguistic patterns and to organize linguistic patterns, which they observe, arriving at generalizations inductively and verifying deductive rules (Saxena et al., 2008). See, for instance, the result of a search of translations of the Turkish word "baş" (head) into Swedish in Figure 4. The students have to find different meanings of the Turkish and the Swedish words and describe how the different meanings correspond in the two languages.

Figure 4. Swedish translations of the Turkish word "baş" (head)

Such exposure to corpora provides students with the chance not only to extract relevant examples of one or the other linguistic structures, but also provides them material for discussion when they find gaps, to verify and extend their hypothesis and to arrive at generalizations (Saxena et al., 2008). One advantage of using corpora in teaching is that instead of learning about linguistic theories in vacuum - which is considered to be a more passive learning method, where facts are fed to students in form of lectures - students have a chance to test theories themselves against real world data represented in corpora. They are able to learn about the theories or concepts themselves, which is a more active learning method. When syntactically annotated corpora, i.e., treebanks, are used by students as part of their learning, distinction between teaching and research is “blurred”, as students, by discovery procedure (thus, research), learn things for themselves. The use of corpora in teaching can, in this way, affect both teachers’ as well as students’ role. This approach is as equally relevant in a classroom set-up as in self-study situations.
References


