A web-based tool for exploring translation equivalents on word and sentence level in multilingual parallel corpora

Leif-Jöran Olsson and Lars Borin
Department of Linguistics, Uppsala University

ETAP–WebTEq är ett webbaserat verktyg utvecklat i parallellkorpusprojektet ETAP. Verktyget möjliggör bekväm navigering och sökning i länkade parallelltexter, primärt i syfte att utvärdera ord- och meningslänkningsmjukvara, men också för att utforska översättningsekvivalens på ord-, fras- och meningsnivå. Det arbetar med XML-taggade två- och flerspråkiga parallellkorpusar, som är meningslänkade och eventuellt också länkade på ordnivå med andra verktyg framtagna i vår parallellkorpusforskning. Vi redovisar de praktiska och teoretiska överväganden som ligger bakom utformingen av ETAP-WebTEq, och illustrerar verktygets användning med hjälp av en ETAP-korpus bestående av parallell tidningstext på sex språk: svenska, finska, polska, serbiska-bosniska-kroatiska, spanska och engelska.

*
*   *

1 Introduction

Parallel texts — ideally consisting of original texts in one language and their translations into another language (but other combinations are also possible; see Borin forthcoming a) — are an excellent source of information about all kinds of relationships obtaining between an original text and its translation. In the Department of Linguistics at Uppsala University, we are pursuing a parallel corpus project, the aims of which are the collection and automatic annotation of multilingual parallel corpora in various domains. This is the ETAP project.[1] Swedish is one of the languages in all the ETAP subcorpora, normally the source language (SL), combined with several other languages, mostly in the role of target languages (TL). The annotations to be made on the ETAP corpora are of two kinds, (1) part-of-speech (POS) tags, i.e., an annotation for each text token (words and punctuation marks), showing its word class and possibly morphological information, and (2) sentence and word alignment, i.e., the establishment of explicit 'links' between equivalent units in the two language versions making up the parallel text.

In the ETAP project, these annotations are seen primarily as the first step in the development of automatic methods for extracting linguistic information from parallel texts for use in a machine translation system, but parallel corpora annotated in this way form a general resource with many other potential uses, e.g., for empirically based linguistic studies, or for bi- and multilingual lexicography. They can also be used for pedagogical purposes in foreign/second language instruction, in translator training, or for the preparation of textbooks, learners' grammars, etc.

We strive to accomplish the aims of the ETAP project by (re)using, as far as possible, existing publicly available software tools, resources and standards. This in turn means that our software development efforts have been directed towards creating a software infrastructure which permits the reuse and combination of diverse NLP resources. To this end, we are developing a flexible storage format, conforming to the XML version of the Corpus Encoding Standard (CES 2000), where resource combination and interaction will be handled by a so-called *blackboard architecture,* i.e., "a framework in which knowledge can be arranged so that it can be distributed and yet shared among a number of cooperating processes" (Walters & Nielsen 1988: 303).

2 The alignment browser: design issues

Thus, the research efforts in the ETAP project are currently concentrated on devising good ways of combining existing NLP resources—such as POS taggers with different tag sets and trained on different text types (Borin 2000), word alignment and POS tagging (Borin forthcoming b), or combining word alignments for more than one language pair (Borin forthcoming c)—in order to enhance the accuracy of both POS tagging and word alignment. Naturally, we wanted to be able to follow up the effects of applying different such combinations on our (fairly large) corpus. Hence, we

[1] ETAP is the acronym of the project title "Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter"' (in English: "Creating and annotating a parallel corpus for the recognition of translation equivalents"). This project is a part of a joint research programme between the universities in Stockholm and Uppsala, "Translation and Interpreting – A Meeting between Languages and Cultures" financed by the Bank of Sweden Tercentenary Foundation (Riksbankens Jubileumsfond); see <http://ww.translation.su.se>.

Erikoiskielet ja käännösteoria. VAKKI:n julkaisut N:o 26. Vaasa 2000, s. 76 - 84.

perceived a need for an alignment browsing and search tool. Generally speaking, any such tool should fulfil at least two demands:

(1) it should give direct access to all important aspects of the underlying representation. Concretely, in the ETAP corpora, all annotation types (POS tags, sentence and word links, etc.) should be accessible. Further, it should allow multilingual browsing, since the ETAP corpora are multilingual, rather than bilingual parallel corpora. This in turn implies that many kinds of alphabets and character representations should be handled.

(2) it should provide an intuitive (for the user) visualisation of the actions it makes possible and the results which these actions engender. For large data sets, this often implies some kind of filtering and zooming facility, in accordance with Shneiderman's (1998:523) "*visual-information-seeking mantra:* Overview first, zoom and filter, then details on demand".

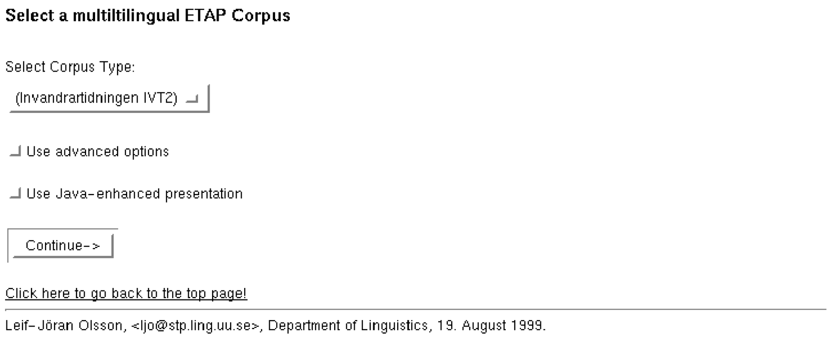
There already exist quite a few alignment browsers, or parallel text concordancers (see, e.g., Kjærsgaard 1986; Barlow 1995; Eberling 1998; Stahl forthcoming), but the ones that we could find in the literature all work with the sentence as the smallest alignment unit, providing general SL and TL word and string search facilities within sentence alignments, and presenting the search results in the form of traditional KWIC concordances, although bilingual instead of monolingual. Since we also wanted to be able to show explicit word alignments, if such are present in the corpus (cf. the first demand above), we could not use any of the existing tools directly. Sticking to the principle of reusing as much existing software as possible, we decided to build upon an existing web-based parallel text concordancer,[2] but adding to it functionality for searching, manipulating and visualising word alignments. We also required general filtering and zooming functions, which as a rule are lacking from existing tools, but which are needed for working effectively with large amounts of corpus material. The adoption of a web-based solution was also a direct consequence of the principle of software reuse; we can use standard web-browsers instead of special client applications: Netscape 4.06 or later, or Internet Explorer 4.0 or later, is needed for the default text-only presentation. The visualisation engine is written in Java, so if it is to be used, Java support should be enabled in the browser.

[2] We used a Uplug application (Tiedemann forthcoming), developed in another parallel corpus project in which our department is a partner, the PLUG project (Sågvall Hein forthcoming).

In the remainder of this paper, we first show some examples of the browser, which we have named ETAP-WebTEq—**ETAP** project **Web**-based browser of **T**ranslation **Eq**uivalents—, at work on the IVT2 ETAP subcorpus. This is a parallel corpus of newswire text in 6 languages, or, rather, 5 language pairs: Swedish (the SL) – Finnish, Polish, Serbian-Bosnian-Croatian, Spanish, and English (the TL's). Finally, we discuss how we plan to develop ETAP-WebTEQ further.

3 A walk-through of ETAP-WebTEQ

3.1 Getting started: selecting a subcorpus



Select a multiltilingual ETAP Corpus

Select Corpus Type:

(Invandrartidningen IVT2)

□ Use advanced options

□ Use Java-enhanced presentation

Continue->

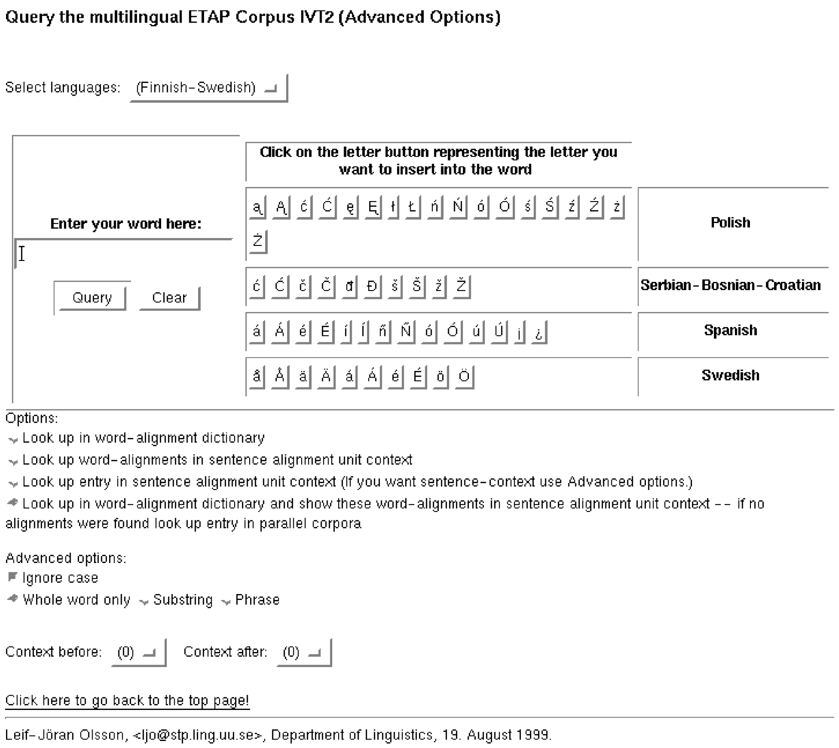
Click here to go back to the top page!

Leif-Jöran Olsson, <ljo@stp.ling.uu.se>, Department of Linguistics, 19. August 1999.

Figure 1: Selecting a subcorpus

After logging in (this is necessary because, for various reasons, not all ETAP subcorpora are freely available), the user is asked to select a subcorpus. In the example shown in Figure 1 the IVT2 corpus has been selected.

3.2 Searching a corpus: Submitting a query



Query the multilingual ETAP Corpus IVT2 (Advanced Options)

Select languages: (Finnish-Swedish)

Enter your word here:

Click on the letter button representing the letter you want to insert into the word	
ą Ą ć Ć ę Ę ł Ł ń Ń ó Ó ś Ś ź Ź ż Ż	**Polish**
ć Ć č Č đ Đ š Š ž Ž	**Serbian-Bosnian-Croatian**
á Á é É í Í ñ Ñ ó Ó ú Ú ¡ ¿	**Spanish**
å Å ä Ä á Á é É ö Ö	**Swedish**

Options:

○ Look up in word-alignment dictionary

○ Look up word-alignments in sentence alignment unit context

○ Look up entry in sentence alignment unit context (If you want sentence-context use Advanced options.)

● Look up in word-alignment dictionary and show these word-alignments in sentence alignment unit context -- if no alignments were found look up entry in parallel corpora

Advanced options:

☑ Ignore case

● Whole word only ○ Substring ○ Phrase

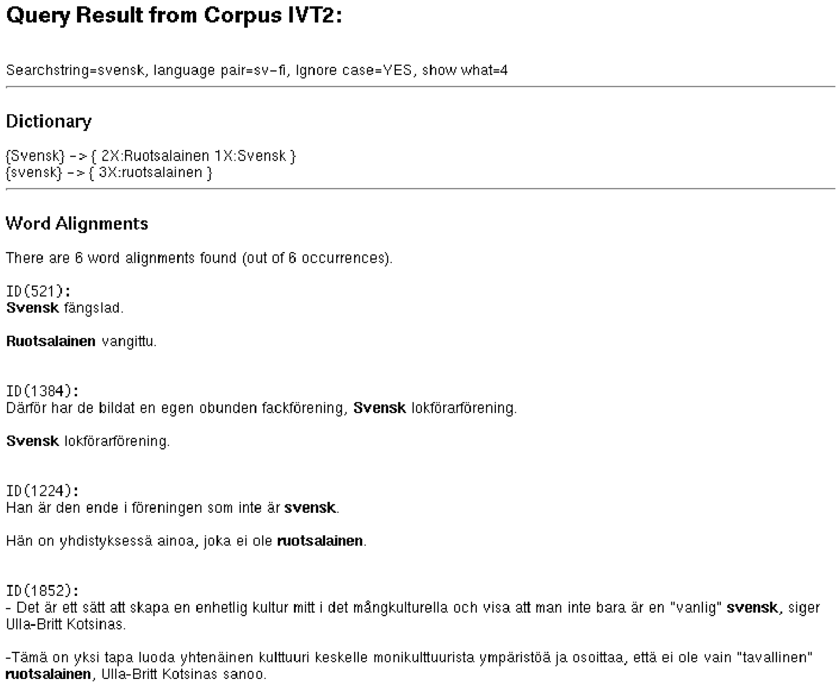
Context before: (0) Context after: (0)

Click here to go back to the top page!

Leif-Jöran Olsson, <ljo@stp.ling.uu.se>, Department of Linguistics, 19. August 1999.

Figure 2: The advanced query form

In Figure 2, you see the various ways in which an aligned corpus can be searched for word alignments. The presentation context can be set to the desired number of sentence alignment units occuring before and after the alignment unit returned as a result of the query. Although the web interface uses the Unicode character encoding, there will still be problems in entering many letters from any standard keyboard. Thus, letters not in the default character encoding standard (ISO 8859–1, i.e. Latin-1, in Figure 2) can be entered by clicking the corresponding letter button.



Query Result from Corpus IVT2:

Searchstring=svensk, language pair=sv-fi, Ignore case=YES, show what=4

Dictionary

{Svensk} -> { 2X:Ruotsalainen 1X:Svensk }
{svensk} -> { 3X:ruotsalainen }

Word Alignments

There are 6 word alignments found (out of 6 occurrences).

ID(521):
Svensk fängslad.

Ruotsalainen vangittu.

ID(1384):
Därför har de bildat en egen obunden fackförening, **Svensk** lokförarförening.

Svensk lokförarförening.

ID(1224):
Han är den ende i föreningen som inte är **svensk**.

Hän on yhdistyksessä ainoa, joka ei ole **ruotsalainen**.

ID(1852):
- Det är ett sätt att skapa en enhetlig kultur mitt i det mångkulturella och visa att man inte bara är en "vanlig" **svensk**, siger Ulla-Britt Kotsinas.

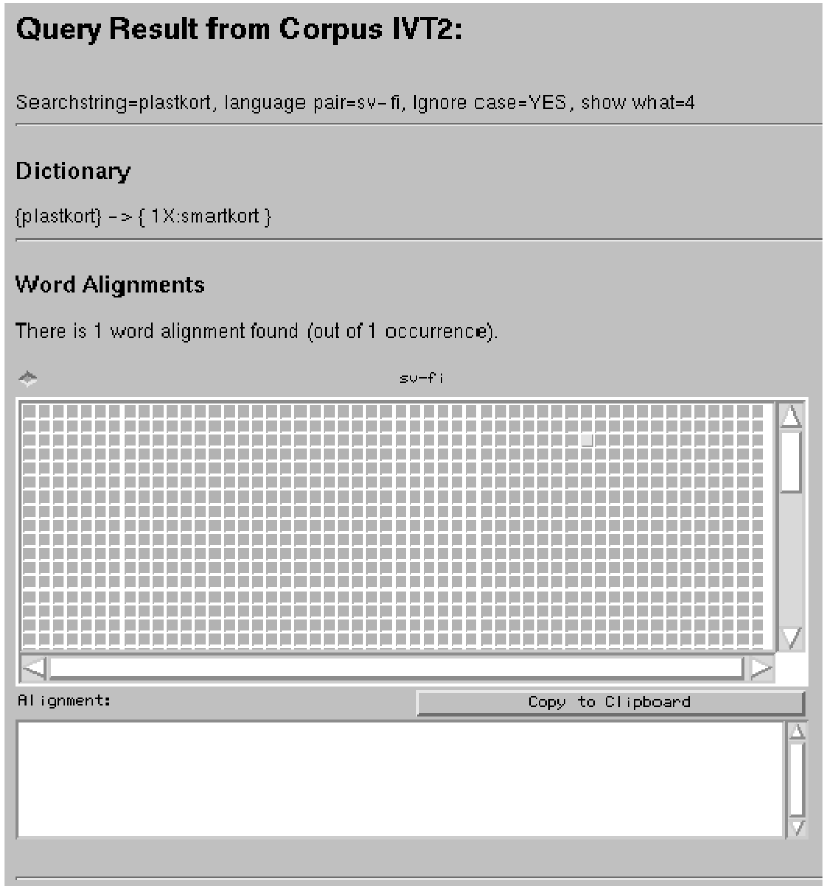
-Tämä on yksi tapa luoda yhtenäinen kulttuuri keskelle monikulttuurista ympäristöä ja osoittaa, että ei ole vain "tavallinen" **ruotsalainen**, Ulla-Britt Kotsinas sanoo.

Figure 3: Search results

In Figure 3, we show the results of a search for the word *svensk* ‘Swedish, Swede’ in the Swedish–Finnish portion of the IVT2 corpus. This corpus has been word aligned, and *svensk* is in the alignment dictionary, together with pointers to all its (aligned) occurences in the corpus, which fact is utilised be the search procedure in preparing the result of the query.

3.3 Visualisation features

The visualization metaphor is under development. To date we have implemented some features having to do with the physical orientation within the corpus.



Query Result from Corpus IVT2:

Searchstring=plastkort, language pair=sv-fi, Ignore case=YES, show what=4

Dictionary

{plastkort} -> { 1X:smartkort }

Word Alignments

There is 1 word alignment found (out of 1 occurrence).

Figure 4: Visualising the distribution of a word alignment in the corpus

In a Java-enabled browser, the user can get an overview of the distribution of the sought word alignment in the corpus. In Figure 4, we show how this is accomplished in ETAP-WebTEq. Each square represents one sentence alignment unit, and those units which contain the word alignment in question are shown in a different colour from the rest (yellow instead of grey; in Figure 4, there is one yellow square, in the third row from the top), and if clicked, show the actual sentence alignment unit, as in the examples in the previous section. This kind of overview is valuable for many reasons, e.g. for finding thematically defined parts of the corpus, but also for isolating systematic failures in the word alignment software.

4 Development plans for ETAP-WebTEq

The present version of ETAP-WebTEq represents a first attempt at a word alignment browser for multilingual parallel corpora. Even if it is already proving its worth as a tool in the ETAP project, there are still many ways in which it could be enhanced, and which we already are or soon will be working to implement and incorporate in the tool:

(1) The alignment overview (see section 3.3, above) should be provided at several successive levels, doing away with the need for scrolling in order to see all alignments. In other words, the degree of detail at each level should be such that all items can be seen simultaneously on the screen, which in practice would mean that at all levels except the lowest, each square would represent more than one sentence alignment unit. This means that numerical information of various sorts must be added, such as the number of alignment units covered by the overview, and where it starts and ends in relation to the whole corpus (this could also be shown graphically).

(2) At present, only the yellow squares in the overview (i.e., those containing the sought word alignment) are clickable. All squares should be expandable in the same way.

(3) In individual sentence alignment units, it should be possible to see which other words and phrases are aligned (i.e., are in the alignment dictionary) in the same unit.

References

Barlow, Michael (1995). ParaConc: A Concordancer for Parallel Texts. In: *Computers & Texts,* Vol. 10, December 1995.

Borin, Lars (2000). Something Borrowed, Something Blue: Rule-Based Combination of POS Taggers. In: *Second International Conference on Language Resources and Evaluation. Proceedings, Volume I.* Athens: ELRA. 21–26.

Borin, Lars (forthcoming a). And Never the Twain Shall Meet? In: L. Borin (ed). *Parallel Corpora, Parallel Worlds. Papers Presented at a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999.* Dept. of Linguistics, Uppsala University.

Borin, Lars (forthcoming b). Alignment and Tagging. In: L. Borin (ed). *Parallel Corpora, Parallel Worlds. Papers Presented at a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Dept. of Linguistics, Uppsala University.

Borin, Lars (forthcoming c). You'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment. In: *COLING 2000 –Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken: DFKI.

CES = Corpus Encoding Standard, <http://www.cs.vassar.edu/CES/>

Eberling, Jarle (1998). The Translation Corpus Explorer: A Browser for Parallel Texts. In: S. Johansson and S. Oksefjell (eds). *Corpora and Cross-linguistic Research. Theory, Method, and Case Studies*. Amsterdam: Rodopi. 101–112.

Kjærsgaard, Poul Søren (1986). REFTEX – et datamatstøttet oversættelsessystem. In: F. Karlsson (ed). *Papers from the Fifth Scandinavian Conference of Computational Linguistics*. Publications No. 15. University of Helsinki, Department of General Linguistics. 121–130.

Sågvall Hein, Anna (forthcoming). The PLUG Project: Parallel Corpora in Linköping, Uppsala, Göteborg: Aims and Achievements. In: L. Borin (ed.). *Parallel Corpora, Parallel Worlds. Papers Presented at a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Dept. of Linguistics, Uppsala University.

Shneiderman, Ben (1998). *Designing the User Interface. Strategies for Effective Human–Computer Interaction*. Reading, Massachusetts: Addison-Wesley.

Stahl, Peter (forthcoming). Building and Processing a Multilingual Corpus of Parallel Texts. In: L. Borin (ed). *Parallel Corpora, Parallel Worlds. Papers Presented at a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Dept. of Linguistics, Uppsala University.

Tiedemann, Jörg (to appear). Uplug – a Modular Corpus Tool for Parallel Corpora. In: L. Borin (ed). *Parallel Corpora, Parallel Worlds. Papers Presented at a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Dept. of Linguistics, Uppsala University.

Walters, John R. and Norman R. Nielsen (1988). *Crafting Knowledge-Based Systems. Expert Systems Made Easy / Realistic*. New York: Wiley.

Lars Borin
Uppsala universitet
Inst. för lingvistik
Box 527
S-751 20 Uppsala
lars.borin@ling.uu.se

Leif-Göran Olsson
Uppsala universitet
Inst. för lingvistik
Box 527
S-751 20 Uppsala
ljo@stp.ling.uu.se