# Interactive Word Alignment for Language Engineering

**Lars Ahrenberg, Magnus Merkel, Michael Petterstedt**
Department of Computer and Information Science
Linköping University
{lah,magme,g-micpe}@ida.liu.se

## Abstract

In this paper we report ongoing work on developing an interactive word alignment environment that will assist a user to quickly produce accurate full-coverage word alignment in bitexts for different language engineering tasks, such as MT lexicons and gold standards for evaluation. The system uses a graphical interface, static and dynamic resources as well as machine learning techniques. We also sketch how the system is being integrated with an automatic word aligner.

## 1 Introduction

Automatic word alignment systems have proved to be useful tools for various language and NLP tasks, such as bilingual lexicon extraction for lexicography, bilingual terminology and machine translation. Although performance is improving, (precision in the range from 80 to 95 per cent, recall slightly above 50%), there are applications, such as translation and the creation of gold standards, where these figures are not good enough. Furthermore, since most automatic systems rely on co-occurrence, rare correspondences go unnoticed, even though they may be relevant for applications such as terminology or lexicography. This means that even for these applications higher recall and precision will give better effect.

For machine translation errors in alignment are likely to cause errors in translation. Thus, either there will have to be a reviewing process when a generated bilingual dictionary is to be part of a MT system, or else the reviewing could be made in the underlying files, i.e., by interactive reviewing. Extending the application area to more linguistic fields, such as translation studies or any form of corpus-based linguistics, errors are of course a curse. Also, observations and generalisations would be better grounded if they are complete, i.e. all instances of the phenomena of interest have been found and classified.

In this paper we present a system and a method to improve the performance of word alignment, by putting a human in the alignment loop. Alignment bears resemblance to translation and, as with translation, systems could improve by learning from human decisions. Kay's argument for the role of humans in translation holds for alignment too; i.e., we should "expect better performance of a system that allows human intervention as opposed to one that will brook no interference until all the damage has been done" (Kay 1997, p 22).

An interactive approach to word alignment requires an efficient interface to review, modify and create alignments. The main features of our system are that it proposes alignments to the user on the basis of its resources and that it is able to improve on its performance by learning from the user sentence-by-sentence.

## 2 Previous work

Most word alignment systems that have been presented to date are automatic, exploring the co-occurrences of terms in large parallel corpora to generate translational equivalences among word types. In addition to co-occurrence data, some systems employ linguistic knowledge of varying levels of sophistication (Melamed 2001, Ahrenberg et al., 2000b, Gaussier et al. 2000). Manual word alignment with the support of interactive tools has been used mainly for the creation of gold standards for evaluation purposes (e.g. Melamed, 2001; Véronis and Langlais, 2000; Ahrenberg et al., 2000a). However, the idea of improving the outcome of an automatic system, though quite common with sentence aligners and the creation of tree-banks (Marcus et al., 1993),

seems not to have been applied systematically to word alignment. Isahara and Haruno (2000) present a post-editing tool for sentence alignment that has been extended with functions for alignment of phrases and proper nouns. In the Cairo system (Smith and Jahr, 2000) a user can examine visualizations of the word alignments produced by a word aligner, but is not allowed to make changes to them.

In Ahrenberg et al. (2002) an earlier version of the interactive linker was presented. This version had a more primitive interface and also lacked several of the resource and learning capabilities included in the current version.

## 3 I*Link – an interactive word aligner

The current version of I*Link supports the following tasks:[1]

- Manual word alignment
- Automatic proposals of token alignments
- Reviewing and editing alignment proposals from the system in an orderly fashion,
- Configuring the resources to be used by the system in a work session
- Compiling reports and statistics from aligned files.

### 3.1 The I*Link workspace

The system has a graphical interface that allows direct manipulation and interaction with static and dynamic resources. The interface is divided into four windows: the *Link Panel*, the *Link Table Panel*, the *Resource Panel* and the *Settings Panel*. In the *Link Panel*, where the current sentence pair is shown, the user can manually select correspondences, or interact with the automatic proposals from the system that can be accepted or rejected according to the user's preferences.

All alignments that are confirmed by the annotator will be marked in corresponding colors in the Link Panel. Furthermore, the alignments are also visualized in a table representation in the *Link Table Panel*. The workspace of I*Link including the Link Panel, Resource panel and Link Table Panel is shown in Figure 1.

[1] Further information about I*Link and downloads can be found at http://www.ida.liu.se/~nlplab/ILink/.

I*Link supports different strategies for how to select and present alignment proposals to the annotator. For example, the annotator can decide that alignments should be presented from left to right based on the source sentence, or that proposals should be given based on the over-all ranking of the alignments that I*Link has made.

### 3.2 Input and Resources

The input to I*Link consists of parallel source and target files which have been aligned on the sentence level beforehand. Input files may be numbered text files or annotated files in XML-format. The annotation records linguistic information on four levels: word form, base form, part-of-speech with morphosyntactic features and syntactic functions, such as subject, object, and attribute, etc. Annotated files generally allow for more sophisticated resources to be used and created. In our project we use the FDG parser from Connexor for linguistic analysis (Tapanainen & Järvinen 1997).

The Resource Panel displays the configuration of active resources for an alignment project. There are basically three types of resources available in the current version: *static resources, dynamic resources* and *patterns*. All types of resources could in principle be used on the four different levels of abstraction supported by the system. Static resources are set up at the start of the alignment project and do not change during the session. Typical examples of static data are bilingual term lists and core lexicons. Dynamic resources on the other hand do change during the session. The third type of resource used in I*Link are pattern resources. These resources define correspondences for tokens such as cognates, numbers and punctuation characters.

### 3.3 Interactive alignment and learning

The learning approach taken in I*Link is based on the fact that the dynamic resources are updated incrementally during the manual revision stage. Each time the user confirms a proposed link the information inherent in the link is stored in the different dynamic resources. The inflected word forms will be added to the word form resources and the base forms to the lemmatized dynamic resources. Also, new information on POS correspondences and syntactic functions

will be put in the dynamic resources. However, if a user rejects a proposal this information is stored as negative data in the dynamic resources on all applicable levels. The updating of the dynamic resources is made incrementally which means that the new information is available immediately for I*Link and can be applied when new proposals are made in the next sentence pair. In our own tests, the improvements from the learning strategies are clearly observable even after a rather limited number of sentence revisions.

### 3.4 Analysis and reports

I*Link contains some additional tools for data analysis. The Link Inspector functions like a fine-grained bilingual concordance program in that it is possible to define search criteria on all combinations of representation levels, word form, base form, POS and function. For example, one can search for all alignments where a subject noun corresponds to an object noun, an adjectival construction corresponds to a verb construction, etc.

There are also inspectors for viewing, searching and editing the static and dynamic resources and a Link Reporter that can summarize and configure the information in the database, including compiling fine-grained concordances according to the user's preferences.

## 4 Different alignments strategies

Word alignment can be used for a number of different applications and tasks, as mentioned in the beginning of this paper. In some applications word form correspondences (Eng: *the soldiers*-Sw: *soldaterna*) are more important; in others only the base forms are to be considered (*soldier-soldat*). This means that decisions have to be made on how to treat articles, prepositions and other function words in the alignment process. For instance, in the examples above, some kind of consistent treatment of articles are called for; when should they be part of an alignment, when should they be treated as deletions, and when should they be ignored. Guidelines that support a specific purpose of word alignment are therefore extremely important to guarantee consistency and valid data across a whole alignment project.

### 4.1 Evaluation

In a small experiment the speed and consistency of four I*Link users were measured. All subjects were familiar with the system though the guidelines to be followed were explained and discussed in a prior session of only twenty minutes. The number of created alignments per subject and minute varied between 12.9 and 16.7 with 14.4 as a mean. 83.4% of the alignments were common to all subjects. If null links were ignored, these showing the greatest variation, agreement rose to 90.4% for the three subjects that were most consistent. Details on this evaluation experiment can be found in Merkel et al. (2003).

## 5 Extensions to I*Link

The current version of I*Link is a stand-alone word alignment tool that proposes token links and interacts with the user. To improve speed we are currently adding a fully automatic mode to the system. The output from the automatic mode can then be reviewed by the user interactively The user will go through a subset of the automatically generated links (for example the first 50 sentence pairs) and then the automatic component will take over again and re-align everything that has not been verified by the user, with the aid of the new information stored in the dynamic resources.

A statistical component for I*Link has been implemented that creates co-occurrence data as static resources on the different description levels. These resources will be utilised by both the interactive and automatic components in I*Link in the next version. Dynamic resources can also be reused across alignment projects depending on the text type.

## References

Ahrenberg L., Merkel, M. Sågvall Hein, A. & J. Tiedemann, 2000a. Evaluating Word Alignment Systems. In *Proc. of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000)*, Athens, Volume III: 1255-1261.

Ahrenberg, L Andersson M, and M. Merkel, 2000b. A knowledge-lite approach to word alignment.. In J. Véronis (ed.) 2000: 97-116.

Ahrenberg, Lars, Magnus Merkel, Mikael Andersson. 2002. A System for Incremental and Interactive Word Linking. In Third International Conference on Language Resources and Evaluation, Las Palmas, 485-490.

Gaussier, E, Hull, D. & S. Aït-Mokhtar, 2000. Term alignment in use - Machine-aided human translation. In J. Véronis (ed.) 2000: 253-276.

Isahara, H. and Haruno, M. 2000. Japanese-English aligned bilingual corpora. In Véronis, J. (ed.) 2000, 313-334.

Kay, M. 1997. The Proper Place of Man and Machine in Language Translation. In *Machine Translation* Volume 12, Nos. 1-2, 1997, 3-23 (reprint from 1980).

Marcus, M., B. Santorini and M. A. Marcinkiewicz, 1993. Building a Large Annotated Corpus for English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

Melamed, I. D., 2001. *Empirical Methods for Exploiting Parallel Texts*. Cambridge, MA, The MIT Press 2001.

Merkel, M., Petterstedt, M. and L Ahrenberg 2003. Interactive Word Alignment for Corpus Linguitics. To appear in the Proceedings of Corpus Linguistics 2003.

Smith, N. A. and M. E. Jahr, 2000. Cairo: An Alignment Visualization Tool. Second Conference on Language Resources and Evaluation, Athens, 2000. Vol I: 549-551.

Tapanainen, P. and T. Järvinen, 1997. A non-projective dependency parser. Proceedings of the 5th Conference on Applied Natural Aslin. E.J., 1949. Photostat recording in library work. In *Aslib Proceedings*, 1:49-52.

Véronis, J. (ed.). 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht, Kluwer Academic Publishers.

Véronis, J. and P Langlais, 2000. Evaluation of parallel text alignment systems. In Véronis, J. (ed.) 2000, 369-388.

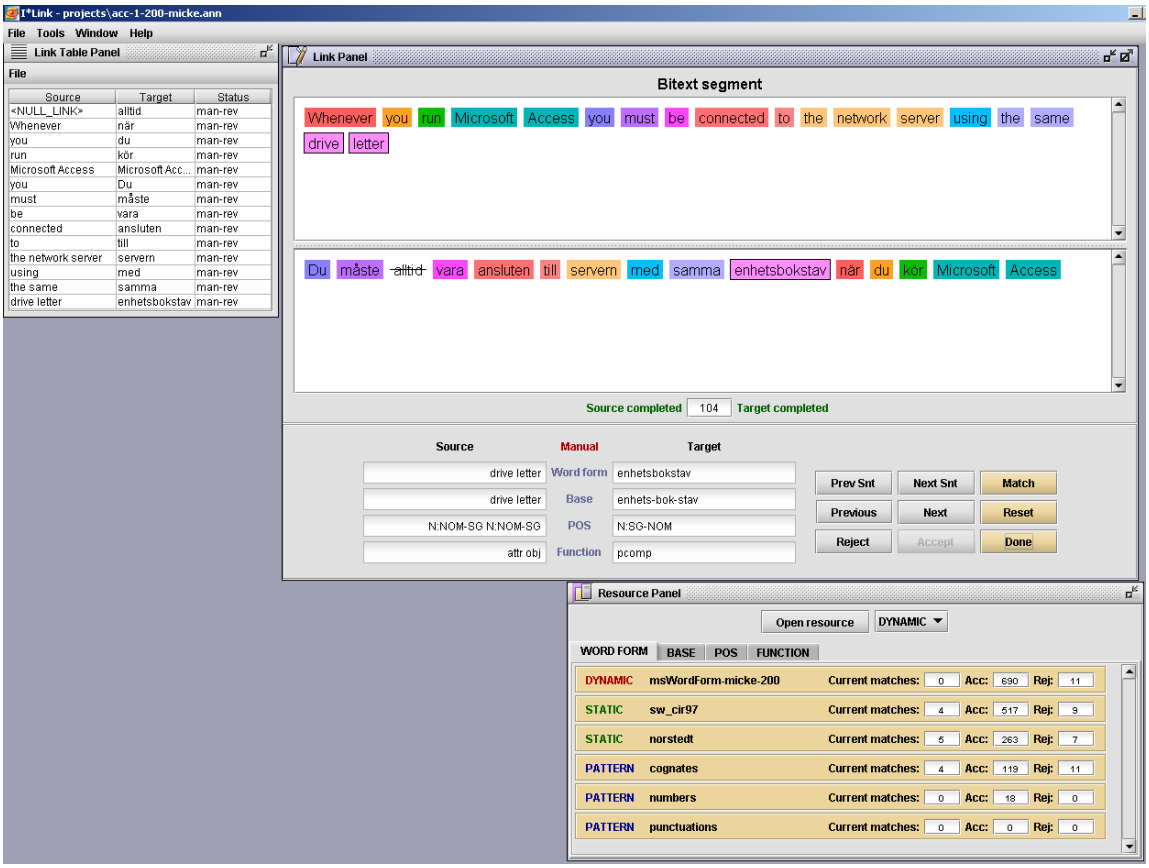


Figure 1. The I*Link interface, including the Link Panel, the Link Table Panel and the Resource Panel. The Link Panel displays the alignments by color-coding. The properties of the alignment in focus are displayed in the lower part of the panel, with values for each description level. To the right of the properties the action buttons are situated. All color-coded alignments are also shown in the Link Table Panel in the upper left corner.