A System for Incremental and Interactive Word Linking

Lars Ahrenberg, Mikael Andersson, Magnus Merkel

Dept. of Computer and Information Science,
Linköping University
S581 83 Linköping, Sweden,
{lah,miand,mme}@ida.liu.se

Abstract

Aligned parallel corpora constitute a critical information resource for a great number of linguistic and technological endeavors. Automatic sentence alignment has reached a level whereby large parallel documents can be fully aligned with the aid of interactive post-editing tools. Word alignment systems have not yet reached the same level of performance, but are good enough to support full word alignment if embedded in an interactive system. In this paper we describe a system for fast and accurate word alignment currently under development at our department, where the user can review and improve the output from an automatic system in an incremental fashion.

1. Introduction

Parallel corpora constitute a critical information resource for a great number of linguistic and technological endeavours, such as contrastive language studies, translation studies, lexicography, terminology, machine translation and cross-linguistic information retrieval.

It has been shown that parallel corpora are useful even in raw form. With minimal preparatory steps such as sentence alignment, they can be built into parallel concordances that can be searched for a number of purposes by the interested linguist. They can also be used for training statistical models of machine translation or for finding word associations.

It is generally true for a corpus, however, that the more information you put into it, the more you are able to get out of it (Leech, 1997). This is certainly true also for parallel corpora. The word alignment system we describe in this paper works on parallel texts that have been annotated with syntactic information and potential multi-word units before word alignment is undertaken. Moreover, the system is interactive allowing the user to provide corrections and additions that the system makes use of in the next iteration. In this way it is possible to build translation databases with very high precision and recall, where grammatical and lexical information is encoded for each link. Also included is information on units of the source that have not been translated and units of the target that have been added. In addition, translation-specific information concerning shifts of various kinds can be registered. We will refer to this result as full word alignment in the sequel.

The paper is organized as follows: In Section 2 we elaborate the background and the motivation for this work. In Section 3, which constitutes the main part of the paper, we describe the system and its architecture. Section 4 provides details about the current status and development of the system.

2. Background

2.1. 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. 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 other NLP tasks such as the creation of tree-banks (Marcus et al., 1993), seems not to have been applied systematically to word alignment. Isahara and Haruno (2000) report on a post-editing tool for sentence alignment that has been extended with functions for alignment of phrases and proper nouns. The Cairo system (Smith and Jahr, 2000) allows the user to examine visualizations of the word alignments produced by a word aligner, but does not allow the user to make changes to them.

2.2. Motivations

Automatic word alignment systems are not yet powerful enough to yield an accurate full alignment. Precision rates are often less than 90 per cent with a recall of about 50%. On the other hand, manual alignment is too slow and expensive. For this reason it seems a good idea to try to combine the abilities of man and computer in such a way that the strengths of both of them can be used to the best advantage. This speaks for an interactive system, where the user can supply the accuracy and the automatic system supply the speed. By using an automatic system that is able to learn from, or directly use the annotations of the user, this arrangement could also mean that the automatic aligner becomes more accurate as work proceeds. This speaks for an incremental work process, where the user and the automatic system take turns in aligning the corpus.

Another disadvantage with automatic aligners is that they only record the existence of links between units in the source and target languages and give no information on the structural and lexical relationship inherent in the links. For this to be possible, however, accurate lexical and syntactic analysis is necessary.

Until very recently a serious problem for smaller languages, such as the Scandinavian languages, has been the lack of general linguistic resources and tools for lexical and syntactic analyses. This forced word alignment tools for such languages to use knowledge-lite approaches with shallow processing, simple modules for string-based

manipulation and a basic statistical approach to word linking (Ahrenberg et al, 2000b). With the recent availability of more general analysis tools for various languages including Swedish, the scene has changed. In this project we use the Functional Dependency Grammar parsers of Conexor Oy (Tapanainen & Järvinen, 1997) for syntactic analyses of both English and Swedish.

2.2.1. Applications

Full high-quality word alignment would be of use for most tasks that are based on parallel corpora. As already mentioned the creation of gold standards for the evaluation of automatic word alignment systems is one of them. While not all evaluations undertaken so far have used full alignments, a gold standard with full alignment can support more varied and more comprehensive evaluations. The gold standard generated with the Blinker tool (Melamed, 2001) is fully aligned.

Translation studies is another area where the benefit would be high. In translation studies the interest is often with complex phenomena that go unnoticed or even upset an automatic word aligner. Similarly, contrastive linguistics has an interest in studying how languages differ in the use of similar constructions, which requires that correspondences at the level of syntactic function can be registered and searched.

Fully aligned translation corpora would obviously also be of importance to machine translation, as they would provide a firmer basis for the generation of both linguistic and statistical data.

2.3. Kinds of alignment

While full word alignment is something to be desired it does not come easily. Even the human expert quite often has difficulties in determining what corresponds to what in a source text and a translation (cf. Kay, 2000). For this reason it is necessary that the alignment process is supported by detailed guidelines.

Word alignment systems can be designed for different objectives, however. One objective, and so far the dominating one, has been to produce lexical data for bilingual dictionaries. In this case the focus is on content words so function words can usually be ignored (unless they are part of multi-word units).

Another purpose for word alignment can be to provide data for machine translation or contrastive studies. In this case the alignment of function words is essential. These two different objectives will affect the content of guidelines and also the working of the automatic word aligner. To illustrate the difference between the two approaches, consider the following example:

ENGLISH: *The football game was played at Wembley.*
SWEDISH: *Fotbollsmatchen spelades på Wembley.*

Here the subject, *the football game,* is translated by the Swedish *fotbollsmatchen.* Most dictionary alignment systems would ignore the English article and only produce the type *football game - fotbollsmatchen* as an entry in the bilingual dictionary. However, when these sentences are aligned on the token level, we would prefer the definite article to be part of the alignment of the subjects (*the football game - fotbollsmatchen*) and also that the passive construction *was played* was coupled to the Swedish *spelades.*

In an earlier project we developed a set of guidelines for our evaluations (Merkel, 1999b). These guidelines, as well as the interactive system, were restricted to the case of aligning randomly selected units from the source half to their corresponding unit in the target, so the guidelines are now in the process of being extended.

It must be pointed out also that guidelines, no matter how detailed, cannot solve all problems. Melamed (2001) reports inter-annotator agreements in the range of 74 to 90% for all words and 87-95% for content words, although his annotators had detailed guidelines and economic incentives to follow them. Our experience so far indicates that performance could be better than that, but it must be kept in mind that some percentage of the data resulting from the system will always be subject to doubt and alternative analyses. For this reason the system gives the user the option of marking a link as uncertain.

3. The system

This section contains a background to the current system, describing the previous automatic word aligner LWA. Then the overall architecture of the incremental and interactive aligner is given followed by a description of the user interface and of the alignment process.

3.1. Background

From 1997 and onwards, the NLP group in Linköping has cooperated with Uppsala University and this work has resulted in the joint alignment system PWA (PLUG Link Word Aligner) which includes Linköping Word Aligner (LWA) and Uppsala Word Aligner (UWA).

3.1.1. LWA

The Linköping Word Aligner (LWA) is an automatic word aligner which takes input in the form of a bitext divided into segments (Ahrenberg et al., 1998; 2000b). The objective of LWA is to find link instances in a bitext and to generate a non-probabilistic translation lexicon from the link instances. The system combines different knowledge-lite approaches to word alignment (i.e., no linguistic resources such as bilingual dictionaries, lemmatisers or POS taggers are used), including surface patterns for morphology and function word lists. Links are established by means of co-occurrence measures, string similarity comparisons and other simple heuristics. There is also a pre-processing stage where a module for multi-word extraction is used to identify possible multi-word units that are treated as tokens in the alignment phase (see Merkel & Andersson, 2000). The system is iterative, repeating the same process of generating translation pairs from the bitext, and then reducing the bitext by removing the pairs that have been found before the next iteration starts (Melamed, 1997, Tiedemann 1997). The algorithm will stop when no more pairs can be generated or when a given number of iterations have been completed. LWA is implemented in Perl with versions for Linux, Sun Solaris and Windows.

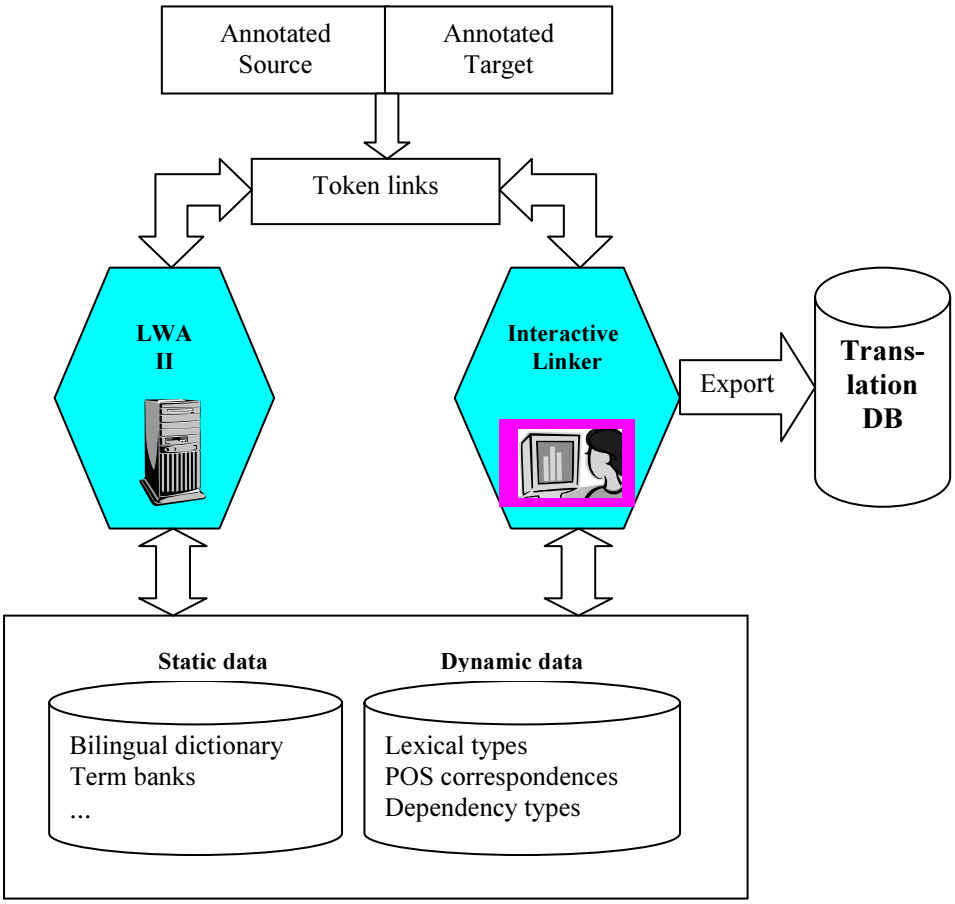


Figure 1. Overall architecture of the incremental and interactive word linking system

3.2. Architecture of the incremental and interactive aligner

A graphical description of the system architecture is given in Figure 1. The two main modules are the automatic word aligner (from now on referred to as LWA II) and the interactive linker which is controlled by a human annotator. The combined system takes three sources as input:

- an xml-annotated source file
- an xml-annotated target file
- a set of token links (initially only sentence links, word links are added at later stages of the process).

The source and target files have first been aligned on the sentence level and then annotated automatically and separately by the Functional Dependency Grammar tools for English and Swedish from Conexor (Tapanainen & Järvinen, 1997). The output from the FDG analyses is further transformed into XML format (including information on base form, syntactic function, parts-of-speech and morphosyntactic features). The kind of information recorded in the monolingual XML file can be illustrated with the following annotation for the English sentence *A bird cried out on the roof*[1].

```
<p id="p1">
    <s id="s1">
        <w id="w1" relpos="1"  base="a"
                func="det" fa="&gt;2" pos="DET"
                msd="SG">A</w>
        <w id="w2" relpos="2"  base="bird"
                func="subj" fa="&gt;3" pos="N"
                msd="NOM+SG">bird</w>
        <w id="w3" relpos="3"  base="cry"
                func="main" fa="&gt;0" pos="V"
                msd="PAST">cried</w>
        <w id="w4" relpos="4"  base="out"
                func="loc" fa="&gt;3" pos="ADV"
                msd="">out</w>
        <w id="w5" relpos="5"  base="on"
                func="ha" fa="&gt;3" pos="PREP"
                msd="">on</w>
        <w id="w6" relpos="6"  base="the"
                func="det" fa="&gt;7" pos="DET"
                msd="SG/PL">the</w>
       <w id="w7" relpos="7"  base="roof"
                func="pcomp" fa="&gt;5" pos="N"
                msd="NOM+SG">roof</w>
</s></p>
```

Figure 2. FDG annotation of a source sentence.

Each word has a unique label (*id*), information of its relative position in the sentence (*relpos*), a functional label (*func*), the base form (*base*), the dependency argument (*fa*), parts-of-speech label and morphosyntactic features (*msd*). All this information is derived automatically from the FDG analysis.

[1] The example is taken from the Linköping Translation Corpus (for more details on the corpus, see Merkel 1999a).

The token links are built up by information on links on different levels which are represented as xml pointers to the source and target file respectively. An extract from the token link file is depicted in Figure 3.

```
<sentLink id="SL1" xtargets="S1;T1">
     <wordLink id="WL1-1" xtargets="w1;w1"
            lexPair="a;en"
            catPair="DET -> DET"
            funcPair="det-> det"
            msdPair="SG -> SG+NOM">
     <wordLink id="WL1-2" method="aut"
            xtargets="w2;w2"
            lexPair="bird;fågel"
            catPair="N -> N" funcPair="subj ->
            subj" msdPair="NOM+SG -> SG+NOM"/>
     <wordLink id="WL1-3 xtargets="w5;w5+w6"
            lexPair="on;uppe+på"
            catPair="PREP -> ADV+PREP"
            funcPair="ha -> ad+advl"/>
    ...
</sentLink>
```

Figure 3. Token links in XML on different levels.

The token links pictured in Figure 3 contain the word link tokens such as *A-En, bird-fågel,* and *on-uppe på* as well as the correspondences on the levels of parts-of-speech, syntactic function and morphosyntactic features.

Apart from utilising the linguistic information contained in the annotated source and target texts, the system also makes use of two different sets of data sources, namely static data sources and dynamic data sources. The static data sources are knowledge sources such as bilingual dictionaries and term banks. The dynamic data sources are the ones that are being built up during the linking process. In the dynamic data sources information on link types that are confirmed by the human annotator are recorded on different levels. By keeping track of all verified links, information on lexical correspondences, parts-of-speech mappings and distributions of syntactic functions in the links, these dynamic data sources will gradually be built up.

3.3. The alignment process

The interactive word linking system can be used in different ways, either in combination with the automatic aligner, but also in a completely manual mode.

The incremental variant of the procedure involves the following steps:

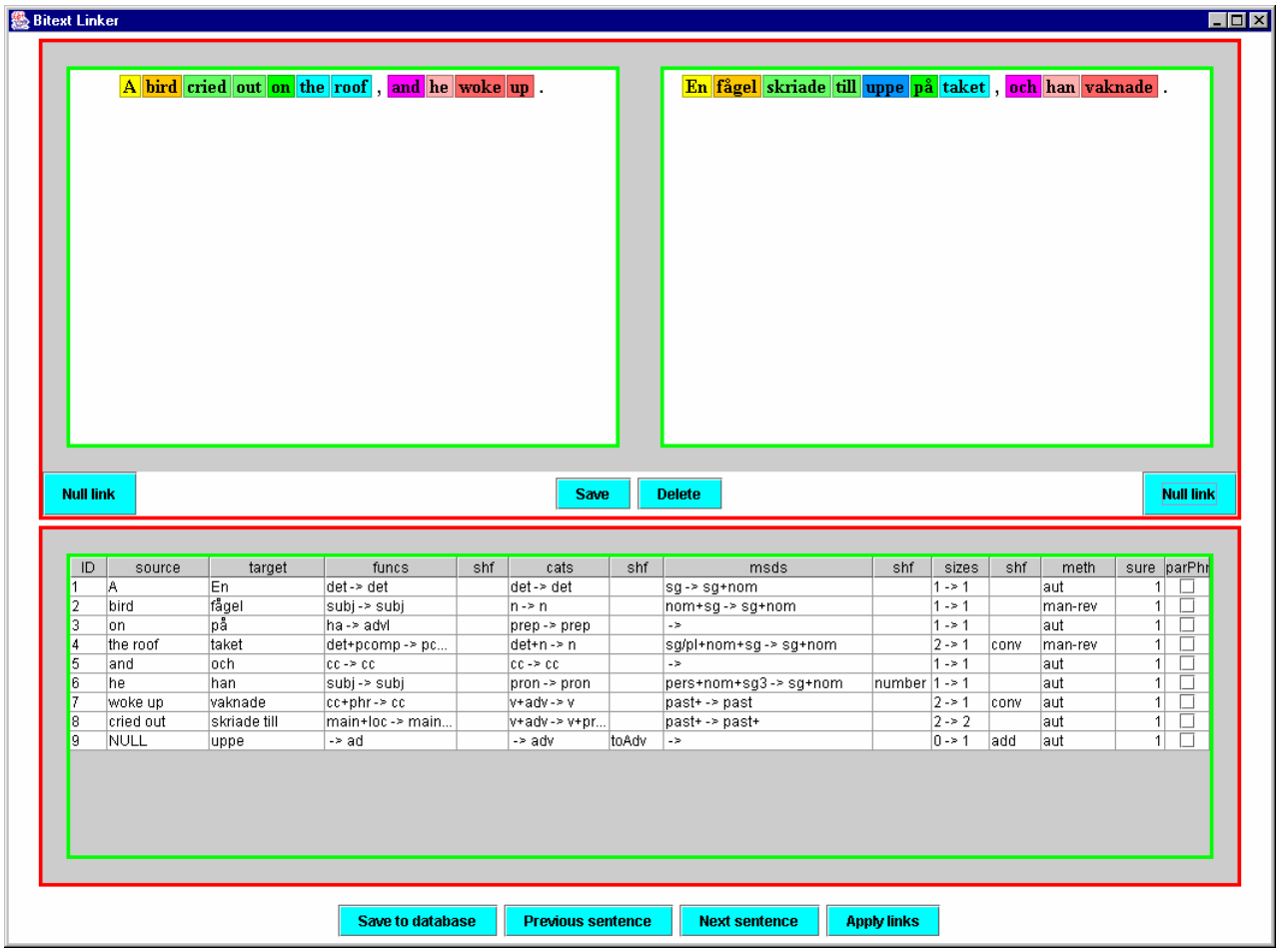
1. The automatic part of the word linker (LWA II) aligns the current parallel corpus on word/phrase level.
2. The user selects an initial set of sentence pairs (e.g. 50 pairs) with all automatically linked word pairs.
3. The user reviews the proposed links, corrects errors and links all tokens that are currently unlinked, with the option of using the data sources available. When the set of sentence pairs are considered correct, the user saves the token link data.
4. The dynamic data sources can now be reviewed by the user and changes can be made.
5. The automatic linking process is resumed on the remaining sentence links. LWA II will now have access to the dynamic data sources which have been created and revised in stage 3 and 4.
6. A new set of sentence pairs are selected and the process resumes from step 3.

And so the process continues until all the sentences pairs have been processed. With more and more information from the dynamic data sources the system performs better and better in the automatic phase, which means less work for the human annotator.

3.4. The user interface

The annotator works in a graphical environment that consists of three panels (see Figure 4):

1. A colour-coded Link Window where source and target units can be selected and linked graphically.
2. A link table including link information such as parts-of-speech shifts (e.g. N-to-A), functional shifts (e.g. subj-to-obj), MSD shifts (present-to-past), link method (automatic or manual), single-word or multi-word link, deletions, additions, convergences, etc.
3. A source and target inspection window where linguistic information from the input files can be inspected by the user (not shown in Figure 4).



ID	source	target	funcs	shf	cats	shf	msds	shf	sizes	shf	meth	sure	parPhr
1	A	En	det -> det		det -> det		sg -> sg+nom		1 -> 1		aut	1	
2	bird	fågel	subj -> subj		n -> n		nom+sg -> sg+nom		1 -> 1		man-rev	1	
3	on	på	ha -> advl		prep -> prep		->		1 -> 1		aut	1	
4	the roof	taket	det+pcomp -> pc...		det+n -> n		sg/pl+nom+sg -> sg+nom		2 -> 1	conv	man-rev	1	
5	and	och	cc -> cc		cc -> cc		->		1 -> 1		aut	1	
6	he	han	subj -> subj		pron -> pron		pers+nom+sg3 -> sg+nom	number	1 -> 1		aut	1	
7	woke up	vaknade	cc+phr -> cc		v+adv -> v		past+ -> past		2 -> 1	conv	aut	1	
8	cried out	skriade till	main+loc -> main...		v+adv -> v+pr...		past+ -> past+		2 -> 2		aut	1	
9	NULL	uppe	-> ad		-> adv	toAdv	->		0 -> 1	add	aut	1	

Figure 4. The user interface for interactive word linking. The token links are coded in identical colours in the interface.

The actual coupling of translation units is first performed in the graphical link window (item 1 above). When a source unit and a target unit have been selected and the linking procedure is applied the information concerning the link is updated in the Link Table. Additions and deletions can also be marked. For some of the translation relations, like POS, functional and MSD shifts, the information is inserted automatically in the Link Table. For more complex relationships, like paraphrases, the annotation is made by the user.

Apart from only revising the automatically produced links, the user can choose to utilise a number of heuristic functions. These functions range from simple similarity identification (cognates) to using the static and dynamic data sources depicted in Figure 1. For example, the system can identify subjects in both the source and target texts as a candidate link, all in the effort to speed up the interactive linking.

4. Discussion and status

The system is written in Java (Java 2, version 1.3) using Swing components for its GUI. The LWA word aligner, written in Perl, is used for the automatic linking phase between each annotation interaction. This integration is at present not implemented in full.

The interactive part is currently running as a stand alone system, allowing the user to use as input ”bare” text files with no links or linked files (as produced by LWA). The new LWA II, that will use linguistic information, is under development.

As described above LWA is a knowledge-lite word alignment system using co-occurrence data as the primary source for deciding translation correspondences. In each iteration the link with the highest score for a given source word (or multi-word unit) is selected, provided it exceeds the set thresholds for frequency and score. The scores can be affected by parameters such as window size, weights and morphological equivalences. The list of candidates can be affected also directly, e.g. if there is a cognate pair on the list of candidates this pair will be moved to the top of the list. When a link type is selected all its instances in the bitext are considered as linked and are not considered in the following iterations.

LWA II is based on the same general principles, but the availability of new types of linguistic data and partially corrected results occasions some changes and improvements.

First of all the fact that the input texts are lemmatized enables the system to base its counts on lemmas rather than word forms. The parts of speech mark-up further allows the disambiguation of common lemmas that are known to introduce errors in the alignment. As the algorithm used is a greedy one, links involving common function words such as *<to, att>* or *<it, det>* tend to overgenerate. By distinguishing occurrences of *to* as infinitive marker and preposition, of *att* as infinitive marker and subjunction, of *det* as pronoun or article etc., these erroneous links can be avoided.

The user's editing of the output from the automatic word aligner yields both positive and negative data. Positive data concern lexical and grammatical correspondences, e.g., parts-of-speech correspondences that have been found in the bitext. Negative data are given by lexical and parts-of-speech correspondences that the user has changed. The user can inspect the type level data generated from his editing and change it before it is given to the automatic system.

The dictionary is used in the system in the same way as the cognate test. If a candidate pair is known to be in the dictionary it is moved to the top of the candidate list provided it satisfies the threshold limits. The lexical correspondences that are generated dynamically are treated in the same way. Also the parts-of-speech correspondences are used to affect the ordering of candidates by eliminating candidates that have a parts-of-speech correspondence judged to be impossible by the user. Future versions of the system may use probabilities associated with parts-of-speech correspondences to affect the score in a more subtle manner.

The user's changes also affect the co-occurrence counts of the automatic system. If a sentence pair has been aligned by the user and signed off as correct, the contribution from that sentence pair to the link counts of a specific source word would fall just on the pair that has been linked and not on any other pair. The more links that are marked as correct for a given word, the more likely will the system be to select translations for that word that it has already encountered in that bitext.

The dependency functions are used in the same manner as the parts-of-speech labels. Thus, they are used only to make categorization of words more fine-grained. It will be interesting to study the behaviour of dependency relations under translation and look for ways to utilize that knowledge for the automatic alignment, but this is future research.

5. Acknowledgements

This work has been supported by The Swedish Research Council within the project ”From parallel corpora to translation databases” and by The Swedish Agency for Innovation Systems within the project ”Corpus-based Machine Translation”. We are also grateful to Michael Petterstedt for discussion and several suggestions on how to improve the system.

6. References

Ahrenberg, L Andersson M, and M. Merkel, 1998. A simple hybrid aligner for generating lexical correspondences from parallel texts. In *Proceedings of COLING-ACL’98,* Montreal, Canada, pp. 29-35.

Ahrenberg L., Merkel, M. Sågvall Hein, A. & J. Tiedemann, 2000a Evaluating Word Alignment Systems. In *Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC-2000)*, Athens, Greece, 31 May - 2 June, 2000, 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.

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 (ed.) 2000: 313-334.

Kay, M. 2000. Preface. In Véronis (ed.) 2000: xv - xx.

Leech, G, 1997. Introducing corpus annotation. In Garside, Leech and McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Pp. 1-18, Longman.

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., 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2),* Providence.

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

Merkel, M, & M. Andersson, 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In the *Proceedings from RIAO 2000,* April 12-14, 2000, Volume 1, pp. 737-746.

Merkel, Magnus, 1999a. *Understanding and enhancing translation by parallel text processing.* PhD Thesis, No. 607, Department of Computer and Information Science, Linköping University.

Merkel, Magnus. 1999b. *Annotation Style Guide for the PLUG Link Annotater.* PLUG report, Linköping University.

Smith, N. A. and M. E. Jahr, 2000. Cairo: An Alignment Visualization Tool. Second Conference on Language Resources and Evaluation, Athens, Greece, 31 May – 2 June, 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.

Tiedemann, J., 1997. Automatical Lexicon Extraction from Aligned Bilingual Corpora. Dimploma thesis, Otto-von-Guericke-University, Magdeburg, Department of Computer Science.

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