Interactive Word Alignment for Corpus Linguistics

Magnus Merkel, Michael Petterstedt & Lars Ahrenberg
Department of Computer and Information Science
Linköpings universitet, S-581 83 Linköping, SWEDEN
Tel. +46 13 28 19 64. Fax: +46 13 28 44 99
{mme,g-micpe,lah}@ida.liu.se

Abstract

In this paper an incremental method and an interactive tool to improve the performance of word alignment are presented. The most important factor in our proposal is to put a human in the alignment loop. This is achieved by using *interactive word alignment*, i.e., a word alignment system and a human that are collaborating in order to make word alignment as efficient and accurate as possible. The aim is full coverage alignments with high accuracy, as the quality of word alignments is crucial in many applications within corpus linguistics, for example, in lexicography, contrastive linguistics and translation studies.

1 Introduction

Automatic word alignment systems are used in various language and NLP tasks, such as bilingual lexicon extraction for lexicography, bilingual terminology and machine translation. Although word alignment has improved, (precision figures are often reported ranging from 80 to 95 per cent), recall is still too low for some more advanced applications. Furthermore, relevant word correspondences may not be discovered for rare words since most automatic systems rely on co-occurrence measures.

As in most areas, alignment errors can have discouraging effects for applications in corpus-based translation studies or any form of corpus-based linguistics. For corpus-based translation studies it is important to be able to identify lexical additions and deletions in the target text, but current automatic word alignment systems lack the capacity to distinguish such operations from cases when the systems simply fail to identify likely source and target correspondences. In other words, a deletion in translation cannot be distinguished from the situation where the system says "sorry, couldn't find a probable alignment". If the alignment of a bitext is complete, i.e. all instances of the phenomena of interest have been found and classified; we would have better justification for generalisations and observations concerning bilingual data. At present there is no automatic word alignment system near achieving complete recall, but by adding a human annotator this would indeed be possible, if appropriate tools were available.

Concordancing tools have been around a long time, primarily for monolingual text, but in later years also for bilingual corpora. One problem with the concordancing tools have been that although they assist the corpus linguist in locating sentences or paragraphs containing the lexical items of interest, there is still the task of locating the exact lexical equivalence pair in the source and target segment. If a lexicographer wants to find out how a source lexical item is rendered in the target text, the concordancing tool will guide her to the sentence/paragraph pairs, highlight the source item, but then she has to pinpoint the target correspondence manually. Given bitexts with correspondences annotated below the sentence level, i.e., correspondences for clauses, phrases and words, a whole new range of possibilities for the corpus linguist would be opened up. Recent corpus applications for monolingual corpora, such as Word Sketch (Kilgarriff & Tugwell 2001) and FrameNet (Fillmore et al. 2002), have shown what can be done given richer linguistic annotation coupled with search and analysis tools. However, to our knowledge this kind of tools has so far only been created for monolingual purposes. One aim of our research is to act as a starting point for creating better bilingual tools for corpus linguistic applications.

In this paper we propose an incremental method to improve the performance of word alignment. The most important factor in our proposal is to put a human in the alignment loop. This is achieved by using *interactive word alignment*, i.e., a word alignment system and a human that are collaborating in order to make word alignment as efficient and accurate as possible. In Ahrenberg et al. (2003), the advantages of interactive word alignment for machine translation systems are discussed in more detail.

An interactive approach to word alignment requires an efficient interface in order to manipulate correspondences between bilingual segments. Such an environment will provide language engineers and corpus linguists with a tool that can help them to quickly produce reference data (gold standards) that can be used to evaluate the performance of their applications. Furthermore, valid reference data will improve the evaluation process of applications for fully automatic word alignment, bilingual lexicon extraction, detection of omissions in translations, etc.

The major innovation of the tool in focus, I*Link, is that there is real interaction between the alignment system and the user. I*Link will propose correspondence candidates, based on information from bilingual resources and built-in heuristics. The user can accept, revise or reject these proposals on the fly. Furthermore, I*Link stores the strategies inherent in the user's choices and adapts its way of suggesting as more and more word alignments are made. This has the effect that the accuracy of the proposed word links is continuously improved during and across word alignment sessions, which in turn means increased efficiency. One important observation regarding interactive word alignment is to acknowledge that there could be several objectives for word alignment. A lexicographer would need correspondence data of a different form than a language engineer involved in developing or tuning a data driven MT system.

In the paper previous approaches to automatic and manual word alignment are described along with some research problems. The system, I*Link, is then described in more detail, including the actual alignment process, the resources used as well as the built-in search and inspection tools. This is followed by a section describing applications within corpus linguistics for parallel corpora and word alignment. The paper ends with a discussion of how an interactive word alignment system could be combined with an automatic system and how such a combined system could be applied within corpus linguistics.

2 Accurate full-coverage word alignment

Let us say that a bitext where all sentence pairs have been assigned an alignment has *full coverage*. Accurate full coverage of a bitext is clearly beyond the capabilities of current automatic word alignment systems. While the methods used for generating probabilistic lexicons and translation probabilities at the word level are very useful, we could clearly generate data with less noise from an accurate full-coverage bitext.

Accuracy in this context must be understood as relative to some set of assumptions and guidelines. Often, as in the two corresponding headings below, there may be differences in opinion as to what an accurate alignment is.

ENGLISH: *They watched the moths in the tobacco flowers.*

SWEDISH: *De följde med blicken fjärilarna nere bland tobaksblommorna.*

(Lit. They followed with the_gaze the_moths down among the_tobacco_flowers.)

For example, one strategy could be to use a construction-oriented approach and align the English article *the* as in the noun phrase alignments of “the moths – fjärilarna” and “the tobacco flowers – tobaksblommorna”. Another strategy could instead leave out the definite articles in the previous example and just couple “moths – fjärilarna” and “tobacco flowers – tobaksblommorna”, if the goal is to compile lexicon-like entries. It could also be discussed whether the verb phrase “följde med blicken” should be regarded as an alignment for the English “watched” and what to do with the unusual translation of “in” to the Swedish“nere bland”.

Even in such simple cases like an English complex verb form *is working* and the corresponding Swedish *arbetar*, where the facts are hardly disputable, one has (at least) two different options: the English copula could be regarded as being deleted or as part of a periphrastic form that is aligned as a unit with the single-form Swedish verb.

If both the source text and the target text of the bitext have been parsed and assigned a linguistic annotation of some sort, the data that can be extracted become even richer (though more complex). In addition, as we show below, the enriched linguistic analysis will make it possible to improve word alignment. In our on-going project, working with English-Swedish parallel texts we use Connexor’s FDG

parsers (Functional Dependency Grammar) for parsing (Tapanainen & Järvinen, 1997)[1]. These parsers provide data of the following kinds for each word token:

- Base form
- Part-of-speech and morphological features
- Syntactic function
- Dependency relation
- Head of dependency relation

The FDG parsers have been developed primarily with monolingual parsing in mind. Thus, their tag sets are not in perfect harmony, but the extra linguistic knowledge provided furthers the actual alignment stage as well as serves as a rich resource when further analysis of the bitext is being performed.

Returning to the previously mentioned objective of full-coverage accurate word alignment there are clear advantages for applications where the word-level correspondences are applied, as in translation studies, contrastive linguistics and lexicography:

- Derived lexical data have a higher quality, and are therefore more reliable
- With a parsed bitext more general and abstract data can be derived at later stages of analysis.

In the following we present the system and its method to improve the performance of word alignment.

3 I*Link – an interactive word alignment system

It has been pointed out before that alignment bears resemblance to translation and, as with translation, systems could improve by learning from human decisions. Martin 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).

In order to review, modify and create alignments with human assistance we need an efficient interface. Anybody who has tried to manually create tables of word correspondences in a word processor or in a spreadsheet, or to draw lines between word tokens in bitext printouts would agree, we think. The main features of our system are that it proposes alignments to the user on the basis of its combined linguistic resources and that it is able to improve on its performance by learning from the user sentence-by-sentence. Furthermore, any errors made can be easily corrected, both on the fly and in post-alignment sessions.

3.1 Previous work

Most word alignment systems to-date have been 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 utilise linguistic knowledge on different levels of sophistication (Melamed 2001, Ahrenberg et al., 2000b, Gaussier et al. 2000, Tufiş 2002). 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.

[1] These parsers are now marketed as Machinese Syntax. See http://www.connexor.com/ for further information.

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 resources and learning capabilities included in the current version.

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

- 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.

The system has a graphical interface that allows direct manipulation and interaction with static and dynamic resources.

3.2 Graphical modules

The graphical 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 alignments of the current sentence pair are presented, the user can manually select correspondences, or interact with the automatic proposals from the system which can be accepted or rejected according to the user's preferences.

All alignments that are confirmed by the annotator will be marked in corresponding colours in the Link Panel. Furthermore, the alignments are also visualized in a table representation in the *Link Table Panel*. The Link Panel is shown in Figure 1 and the Link Table in Figure 2.

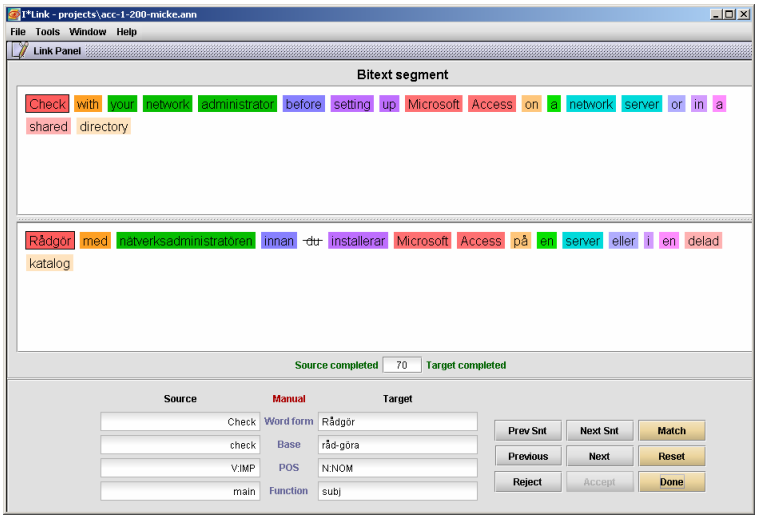
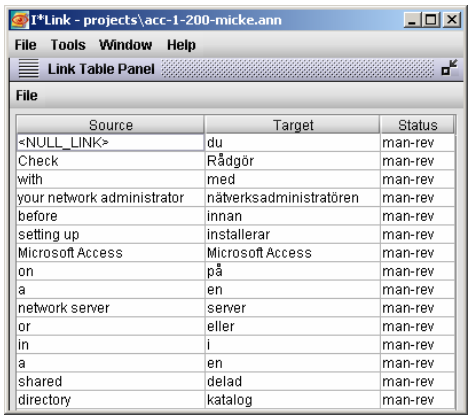


Figure 1. The Link Panel from I*Link. The buttons in the lower right corner allows the user to accept or reject alignments proposed from the system. Alignments are shown in corresponding colours. Additions and deletions are visualized through strike-through lines in the target and source texts.

Apart from annotating straight-forward correspondences, it is possible to represent deletions and additions. The linguistic information for lexical items in focus is visualised in the interface, both in pop-up boxes and directly in the lower part of the Link Panel.



Source	Target	Status
<NULL_LINK>	du	man-rev
Check	Rådgör	man-rev
with	med	man-rev
your network administrator	nätverksadministratören	man-rev
before	innan	man-rev
setting up	installerar	man-rev
Microsoft Access	Microsoft Access	man-rev
on	på	man-rev
a	en	man-rev
network server	server	man-rev
or	eller	man-rev
in	i	man-rev
a	en	man-rev
shared	delad	man-rev
directory	katalog	man-rev

Figure 2. The Link Table showing the same alignments as the Link Panel in a table format.

3.3 Input and Resources

The input to I*Link consists of parallel source and target files which have been aligned on the sentence level in advance. Input files may be line 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 dependency relations, such as subject, object, and attribute, etc.

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 of I*Link, namely, *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. Recurring POS correspondences can also be used.

Dynamic resources change during the session. When the annotator accepts or rejects a proposal from the system, or defines an alignment manually, the action is recorded in the dynamic resources (for example, both as a word form and a base word correspondence, as well as data on the parts-of-speech and syntactic function correspondences). The dynamic resources will therefore contain both positive (accepted and manually defined) data and negative data (rejected proposals). The third type of resource used in I*Link are pattern resources, which define correspondences for tokens such as cognates, numbers and punctuation characters. All resources have the capacity to store positive and negative resources. An example is shown in Figure 3.

```
…
formerly#tidig#1#0
for-example#exempelvis#1#0
for-example#till-exempel#6#0
from#från#14#0
from#i#3#1
from#på#1#0
from#via#1#0
from#<NULL_LINK>#2#0
future#framtid#1#0
gain#<NULL_LINK>#1#0
generate#läsa av#0#1
…
```

Figure 3. Example of a dynamic base form resource. The first field (before "#") holds the source item; the second field contains the target item, and the third number of observed positive instances (acceptances). Finally the number of negative observations (rejection of system proposals) is given. In the given sample the pair “from –i” has been accepted three times and rejected once whereas “generate – läsa av” has no positive observations, only one rejection.

The Resource Panel used for controlling the resources in an alignment project is shown in Figure 4 below.

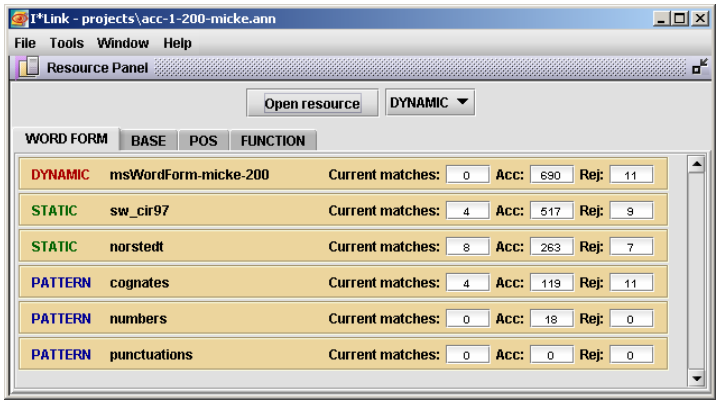


Figure 4. The Resource Panel. On the active word form level six resources are active; one dynamic resource, two static resources and three pattern resources. The figures on the right hand side display the number of current matches in the active bitext segment as well as the total number of acceptances and rejections made by the annotator.

3.4 Heuristics and settings

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 in the order that reflects the over-all ranking, made by I*Link.

3.5 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 their dynamic resources. Also, new information on correspondences for parts-of-speech and dependency relations will be put in the dynamic resources. This also applies when the user adds new alignments manually by selecting items in the Link Panel, making such alignments to be stored as positive data. 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.

3.6 Analysis and reports

To be able to analyse the alignment data, I*Link contains some additional tools. One such tool is the Link Inspector, which 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. This means, for example, that one could identify all the links 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. Examples of how these tools can be used are shown in section 4.

3.7 Performance

In an experimental session we measured the speed and consistency of four subjects. A small set of guidelines was used. All subjects were familiar with the system, but only two of them, A and C, with the guidelines. The guidelines were explained and discussed in a prior session lasting for twenty minutes.

Each user aligned 97 sentence pairs from the help files of Microsoft Access 2000 for XP. The results are shown in Tables 1 and 2.

Table 1 shows results for individuals and speed. The difference between subject B and the other subjects can on the whole be attributed to the use of a different computer, where there is no delay in going from one sentence pair to the next. On the other machines this delay varies between 1-10 seconds depending on the length of the sentences in the pair.

Table 1. Working speed with I*Link for four different users familiar with the system.

Id	No. Sent. Pairs	Total No. Links	Time (min.)	Links per min.	Min. per sent.
A	97	1564	121	12.9	1.25
B	97	1542	92	16.7	0.95
C	97	1551	111	14.0	1.14
D	97	1542	108	14.3	1.11
Mean	97	1550	108	14.4	1.11

Table 2 shows agreement in word alignment between subjects. All subjects agreed on 83.4% of the links produced, and if null links were removed, where alignment strategies were the most varied, the percentage of agreement is as high as 86.5%. Subject D generally was less consistent with the others. If a training session is included so that subjects can discuss disagreements, and guidelines be more detailed, it is likely that these figures can be raised substantially.

Table 2. Agreement in word alignment between subjects.

Common Links (%)	ABCD	ABC	BC	AD
Incl. Null links	83,4	88,7	91,6	85,4
Excl. null links	86,5	90,4	93,1	88,9

The method that the system used for alignment proposals in this experiment was “one-by-one”, i.e., the subjects were given one suggestion at a time, and was asked to give a judgement of this as “accept”, “reject” or ask for a new proposal. An alternative way of presentation would be to ask the system to show all the alignment proposals for a sentence pair immediately and let the user modify the ones that are erroneous. When sentence pairs are short and the proposed alignments have high accuracy, this may speed up the interaction substantially. We have not tested the system with the latter method yet, but we plan to do so.

4 Corpus Linguistics applications for interactive word alignment

As mentioned in the introduction there are interesting applications given full coverage word alignment with high accuracy. For example, for commercial lexicographical work, Atkins (2002, p. 13) makes the following remark concerning bilingual concordancing software: ”A smarter program which would tailor the output of the concordancing program to our needs might persuade reference publishers to change their minds about the use of parallel corpora /…/ Lexicographers need some bilingual form of the Word Sketch tool to help them use parallel corpus data within the time constraints of commercial dictionary production.” Although the inspection and analysis tools included in I*Link are rather limited, they are still very powerful for producing overviews of lexical correspondences in parallel corpora. Consider the brief examples shown in Table 3 and 4 below. Here searches have been made for two kinds of functional shifts in an English-Swedish bitext, namely when source objects have been rendered as target subjects and vice versa.

Table 3. Output from Link Reporter on object-to-subject shifts

Source base	Target base	Source example	Target example	Count (S:2)
mweta	mweta	... one stage , for getting **[Mweta]** banished to the far Western ...	... sin tid styrt om att **[Mweta]** hade förvisats till en avkrok ...	1
you	du	"They expect **[you]** back , " she said ...	"De räknar med att **[du]** ska komma tillbaka " , ...	1

Table 4. . Output from Link Reporter on subject-to-object shifts

Source	Target base	Source example	Target example	Count (S:10)
decision	beslut	... of practical matters by which **[decision]** is broken up into reality ...	... de praktiska problem som förvandlar **[beslut]** till verklighet .	1
he	han	... of dill ; " There **[he]** is , " she said ...	... . " Där har vi **[honom]** " , sade hon .	1
it	den	... code so deeply accepted that **[it]** had never been discussed	... att de aldrig hade diskuterat **[den]**: man stod till förfogande ...	1
she	hon	Because **[she]** was suddenly realizing that it ...	Ty det slog **[henne]** plötsligt att så hade det ...	1
that	som	... bought , filled with possessions **[that]** had been stored all the ...	... köpt och fyllt med tillhörigheter **[som]** de haft magasinerade under alla ...	1
that	som	... true sense of after all **[that]** had gone before) an ...	... (bokstavligen , trots allt **[som]** livet fört med sig) ...	1

/…/

If the lexicographer wants to find examples of when single verbs have been translated with more complex verb phrases, it is possible to express this in the search window and end up with something similar to the examples in Table 5. Here single verbs in the English source text have Swedish correspondences that consist of more complex verb phrases in the form of verb – preposition – noun.

Table 5. Output from the Link Reporter showing single verbs (POS=V) in the source text that correspond to complex verb phrases (V PREP N) in the target.

Source base	Target base	Source POS	Target POS	Source example	Target example	Count (S:3)
garden	påta i träd-gård	V	V PREP N	He and Olivia **[gardened]** on summer evenings , not ...	Han och Olivia brukade **[påta i trädgården]** om sommarkvällarna , inte på ...	1
make	göra i ordning	V	V PREP N	She used to **[make]** packages of sandwiches for Mweta ...	Hon brukade **[göra i ordning]** smörgåspaket åt Mweta att ta ...	1
settle	komma i ordning	EN	V PREP N	... - you 'll be more-or-less **[settled]** by the time she arrives ...	... fall har du mer-eller-mindre hunnit **[komma i ordning** när hon kommer ner	1

Another application for the high-quality word alignment that we are aiming for is to try to derive wordnet relations such as synonymy and hyponymy from translation corpora using the concept of *semantic mirrors* suggested by Dyvik (2002). The idea is to use word correspondences from a parallel corpus to build synonymy and hyponymy sets via the word alignments created by a tool like I*Link. This approach requires data with high accuracy and can therefore be suited to the interactive word alignment approach suggested here. The resulting data sets covering synonymy, hyponymy and semantic nearness could also serve as resources in future versions of I*Link, as a kind of extended semi-semantic lexical resources which could increase the performance of word alignment.

5 Integrating automatic and interactive alignment

The current version of I*Link is a stand-alone word alignment tool that proposes token alignments and interacts with the user. We are currently adding a fully automatic mode to the system where I*Link aligns the bitext without interaction with the annotator. The output from the automatic alignment is then reviewed by the user (by accepting, rejecting, adding and modifying word links). 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.

The automatic alignment component is still at a preliminary stage, but the goal is to make use of the resources available to I*Link, that is, information on word forms, base forms, parts-of-speech and syntactic functions. In addition to these, the automatic component utilises information on co-occurrence statistics for lexical correspondences along the lines of our previous automatic aligner, Linköping Word Aligner, LWA, (Ahrenberg et al. 2000). The main difference between LWA and the current automatic aligner lies in the possibility of exploiting more linguistic information and the approach that the token alignments are in focus, that is, all possible word alignments in one sentence pair is aligned at a time. In evaluations of LWA we found that it had problems with over-generalizations of certain alignments (mostly homographs, such as *att* which can function both as a subjunction and infintive marker in Swedish). Furthermore, multi-word units (MWUs) had a much higher error rate than single word units which was due to the built-in approach in LWA to generalize type alignments to token alignments. In the new approach, the dynamic and static resources are used as global knowledge bases for the whole alignment process, but contextual factors present in specific sentence pairs function jointly with the global resources in a way that steers the alignment to better performance.

6 Future work

As a tool for creating accurate full-coverage word alignment, the current version of I*Link has certain limitations that we intend to overcome in future versions. These extensions will include

- handling of discontinuous multi-word units,
- integration between the interactive mode of alignment and a fully automatic component,
- support for optional user annotations of alignments,
- database connections,
- integration of various support tools for sentence alignment that adhere to the input XML formats,
- improved search and analysis tools,
- support for more syntactic annotations, i.e., other taggers.

7 Concluding remarks

In this paper we have argued that interactive systems is the best way to improve word alignment for corpus linguistics applications and have presented the I*Link system as a first step forward in this direction.

Information about I*Link is continuously being updated at http://www.ida.liu.se/~nlplab/ILink/. Versions for academic use can be downloaded from this site.

References

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

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

Ahrenberg, L., M. Merkel & M. Petterstedt, 2003. Interactive word alignment for language engineering. To be published in the Proceedings of EACL-2003, Budapest.

Atkins, Sue. 2002. Then and Now: Competence and Performance in 35 Years of Lexicography. In Proceedings of the Tenth EURALEX International Congress, Volume 1: 1-28, Copenhagen.

Dyvik, H. 2002. Translations as semantic mirrors: from parallel corpora to wordnet. Paper presented at The 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English (ICAME 23), Gothenburg, 22-26 May, 2002. http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/ICAMEpaper.pdf.

Fillmore, C.J., C.F. Baker, and H. Sato. 2002: The FrameNet Database and Software Tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC). Las Palmas: 1157-1160.

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

Isahara, H. and M. Haruno. 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).

Kilgarriff A, & D. Tugwell. 2001. Word Sketch: Extraction and Display of Significant Collocations for Lexicography. ACL Workshop on Collocation, Toulouse, July 7, 2001.

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.

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 *5th Conference on Applied Natural Language Processing* (ANLP'97): 64-71.

Tufiş, Dan, and A.-M. Barbu. 2002. Lexical token alignment: experiments, results and applications. In Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas: 458-465.

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