Evaluation of LWA and UWA

PLUG Deliverable 3A.1

Lars Ahrenberg
Department of Computer Science
Linköping University
lah@ida.liu.se

Magnus Merkel
Department of Computer Science
Linköping University
magme@ida.liu.se

Anna Sågvall Hein
Department of Linguistics
Uppsala University
anna@ling.uu.se

Jörg Tiedemann
Department of Linguistics
Uppsala University
joerg@stp.ling.uu.se

Abstract
This report describes our work on evaluating the two word alignment systems used as input to the PLUG project, the Linköping Word Aligner (LWA) and the Uppsala Word Aligner (UWA). First we present the chosen evaluation method and give brief overviews of the two systems. In the final sections we discuss the results and draw conclusions for the improvements of the two systems and for the design of a public system based on LWA and UWA.

1 Introduction

This report describes our work on evaluating the two word alignment systems used as input to the PLUG project, the Linköping Word Aligner (Ahrenberg. Andersson & Merkel 1998; henceforth LWA) and the Uppsala Word Aligner (Tiedemann 1998; henceforth UWA). In this section we first present the data and evaluation method used for the experiment. The second section gives brief overviews of the two systems and in the following section their performance on the test data is reported and compared. In the final sections we discuss the results and the evaluation method chosen and conclude by drawing some conclusions for the improvement of the two systems and for the design of a common research system based on LWA and UWA.

A word alignment system is a system that attempts to identify corresponding words and phrases in a parallel text. Both LWA and UWA assume that their input texts have been divided into smaller segments of corresponding text units (sentences or paragraphs) before word alignment starts. We refer to such segments as *bitext segments*, the term *bitext* being used as synonymous with parallel text. A word correspondence in a bitext segment will often be referred to as a *link instance*, or simply as a *link*. A link instance is a pair of *link units*, i.e. single word or multi-word tokens on either side of the bitext, that have been found to correspond. A *link type*, on the other hand, is defined as a pair of corresponding words or phrases that are instantiated somewhere in the bitext. It is quite possible for a link type to have several instances in a single bitext segment This means that a system may be correct at the type level, but link the wrong instances, as in the following example:

```
The crowd was dancing, shouting and waving flags.
Människorna dansade och skrek och viftade med flaggor.
```

1.1 Method and test data

Merkel & Ahrenberg (1999) gives a review of evaluation methods for word alignment systems. The basic conclusion from this review was that evaluations based on aligned reference data, commonly referred to as gold standards, are superior to methods that merely inspect system output, since they rely on correspondence criteria that are independent of any specific system and thus provide a standard for comparison of different systems.

We must of course recognize that word alignment systems can be built for different purposes. A common purpose of word alignment systems is that of finding correspondences that go beyond existing entries in a given bilingual dictionary or term bank. For any such application the choice of test data and method should reflect the purpose at hand. Arguably, however, alignment of word and phrase tokens form the basis of most alignment systems, regardless of purpose, and thus comparison with a gold standard of link instances can be generally applied, at least as one of several aspects of an evaluation.

The basic setup of the evaluations was to create gold standards for a majority of sub-corpora of the PLUG Corpus[1]. The PLUG Corpus consists of parallel texts of different language pairs and genres. The size of sub-corpora varies between 8000 and 340,000 words. The sub-corpora used for the evaluation are shown in Table 1.

[1] An overview of the PLUG corpus can be found at the PLUG home page (http://stp.ling.uu.se/~corpora/plug/).

Table 1. PLUG sub-corpora used for evaluation. The first four letters of the name indicate the language pair, while the fifth letter indicates text type. Size is measured as the number of words in the whole sub-corpus.

Sub-corpus	Language pair	Text type	Size	Systems tested
sventscan	Swedish/English	Technical text	385,000	Both
svdetscan	Swedish/German	Technical text	337,000	UWA
ensvtxl	English/Swedish	Technical text	125,000	LWA
ensvtacc	English/Swedish	Technical text	163,000	Both
svdeprf	Swedish/German	Political text	7,800	UWA
svenprf	Swedish/English	Political text	8,000	UWA
svdepeu	Swedish/German	Political text	180,000	UWA
svenpeu	Swedish/English	Political text	186,000	Both
ensvfbell	English/Swedish	Fiction	132,000	Both
ensvfgord	English/Swedish	Fiction	169,000	LWA

The gold standards were created by randomly generating 500 tokens occurring in different sentences from the source half of each sub-corpus.[2] These were assigned corresponding tokens by human annotators according to a detailed set of guidelines (Merkel, 1999). The two most basic guidelines were the same as those used in the ARCADE project (Véronis and Langlais, forthcoming), namely

- As many tokens as are required to obtain an equivalence should be included in a correspondence;
- No more tokens than are required to obtain an equivalence should be included in a correspondence.

This has the effect that where one language uses syntactic means to express a certain feature and the other language uses morphological means, a multi-word expression will be linked to a single word. A common case is the English definite article corresponding to a definite suffix in Swedish yielding pairs such as *the car : bilen*, in the pair below:

```
John jumped into the car.
John hoppade in i bilen.
```

Other common cases of English multi-word tokens corresponding to Swedish single tokens are compounds such as Eng. ***railway accident*** to Sw. ***tågolycka***, and genitives such as Eng. *the countries* ***of Europe*** to Sw. ***Europas*** *länder*.

Unlike the ARCADE project, we did not try to select tokens on the basis of specified features, but used a randomized process. Thus, the reference data include correspondences between function words as well as content words and phrases consisting of both types of words. In the final chapter we will discuss the advantages and disadvantages with this approach.

[2] For the smaller sub-corpora such as svenprf and svdeprf, only 100 tokens were sampled.

The creation of the gold standard was performed by means of the PLUG Link Annotator Tool (Merkel, Andersson & Ahrenberg, forthcoming). The PLA allows for the annotation of null links (words or phrases that have not been translated) and for categorising links into clear and fuzzy links.

The work was shared between Uppsala and Linköping. In Uppsala the creation of the gold standard was performed by two students working together, while in Linköping two annotators worked independently, one of which had also been involved in the specification of the annotation guidelines. It turned out that the two annotators in Linköping agreed completely on 92% of the cases with a low at 89.8% for svenpeu and a maximum at 95.4% for ensvtxl. The majority of mismatches were due to differences in the interpretation of boundaries of phrasal correspondences. We did not attempt to form a common annotation before the evaluation started. Instead, the links provided by the most experienced annotator have been used.

Altogether 4200 corresponding pairs were obtained.

1.2 Measures

The performance of the systems has been assessed by means of the following measures:

> *Recall* = the proportion of reference links, including null links, that have been retrieved.
>
> *Precision I* = the proportion of retrieved links that are correct or partially correct.
>
> *Precision II* = the same as precision I, except that partially correct links are counted only as 50% correct.
>
> *F-measure* = the geometric mean of recall and precision II.

Recall and precision are standard measures for tasks such as word alignment. However, given that the sample contains null correspondences, i.e., tokens that have no corresponding tokens at the other end, and phrases that may be only partially linked, their definition is not so straightforward. For the purpose of this evaluation we divided the outcomes into four different categories:

> **C**(orrect) = the retrieved link agrees completely with the reference link; this also includes the case where nothing has been proposed for a null link.
>
> **P**(artial) = the retrieved link overlaps with the reference link on both halves, but is not identical.
>
> **I**(ncorrect) = a link has been retrieved that assigns some unit to the source half of a reference link, but this unit has no overlap with the target half of the reference link.

M(issed) = a link in the reference data has been missed completely by the system; this also includes the case where something has been proposed for a null link.

Using these categories, we calculated recall and precision as follows n(X) means the number of occurrences of category X:

Recall = (n(C) + n(P) + n(I)) / (n(C) + n(P) + n(I) + n(M))

Precision I = (n(C) + n(P)) / (n(C) + n(P) + n(I))

Precision II = (n(C) + 0.5n(P)) / (n(C) + n(P) + n(I))

In addition to these measures, we also used other measures for more detailed analyses of the results. These will be explained as they are introduced in the analysis.

The scoring of recall and precision measures was done automatically, though by different modules in Uppsala and Linköping.

2 The two word alignment systems

In this section we give an overview of the Linköping Word Aligner (LWA) and the Uppsala Word aligner (UWA).

2.1 Linköping Word Aligner (LWA)

Linköping Word Aligner has been in operation since the fall of 1997. The version used in this evaluation follows the original design, but has been improved in several respects during the PLUG project in comparison with the version used in Ahrenberg et al., 1998.

The objective for LWA is to find link instances in a bitext and generate a non-probabilistic translation lexicon from them. The system provides output of both kinds.

The system takes input in the form of a bitext divided into segments. The current version requires the bitext segments to be numbered and the same numbers to be used as references on both halves of the bitext.

The system is implemented in Perl with versions for Windows and Sun Solaris.

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). The algorithm will stop when no more pairs can be generated, or when a given number of iterations have been completed. The system maintains a distinction between open class and closed class expressions. The closed class expressions have to be listed by the user.

In each iteration, the following operations are performed:

For each open class expression in the source half of the bitext (with frequency higher than a set value), the open class expressions in corresponding sentences of the other half are ranked according to their likelihood as translations of the given source expression. The ranking is based on statistical word association scores such as the t-score, mutual information or the Dice coefficient. In this evaluation the t-score was used. The target candidate giving the highest score is selected as a translation provided the following two conditions are met: (a) the score is higher than a given threshold, and (b) the overall frequency of the pair is sufficiently high. These are the same conditions that were used by Fung and Church (1994).

This operation yields a list of translation pairs involving open class expressions.

The same as in (i) but this time with the closed class expressions. A difference from the previous stage is that only target candidates of the proper categories for the source expression are considered.

Open class expressions that constitute a sentence on their own (not counting irrelevant word tokens) generate translation pairs with the open class expressions of the corresponding sentence.

When all (relevant) source expressions have been tried in this manner, a number of translation pairs have been obtained that are entered in the output table and then removed from the bitext. This will cause fewer candidate pairs to be considered in the sequel and affect scores by reducing marginal frequencies and changing the contents of link windows. The reduced bitext is input for the next iteration.

The basic mode of operation can be enhanced by a number of options. An overview of the system is given below in Figure 1. The core of the system contains the alignment kernel that uses the word association score machinery to execute the basic processes of word alignment. In addition there are four main modules that can be invoked to improve the performance of the system. Apart from the main modules there are a number of parameters that can be set to determine what options should be used for a particular execution of the program.

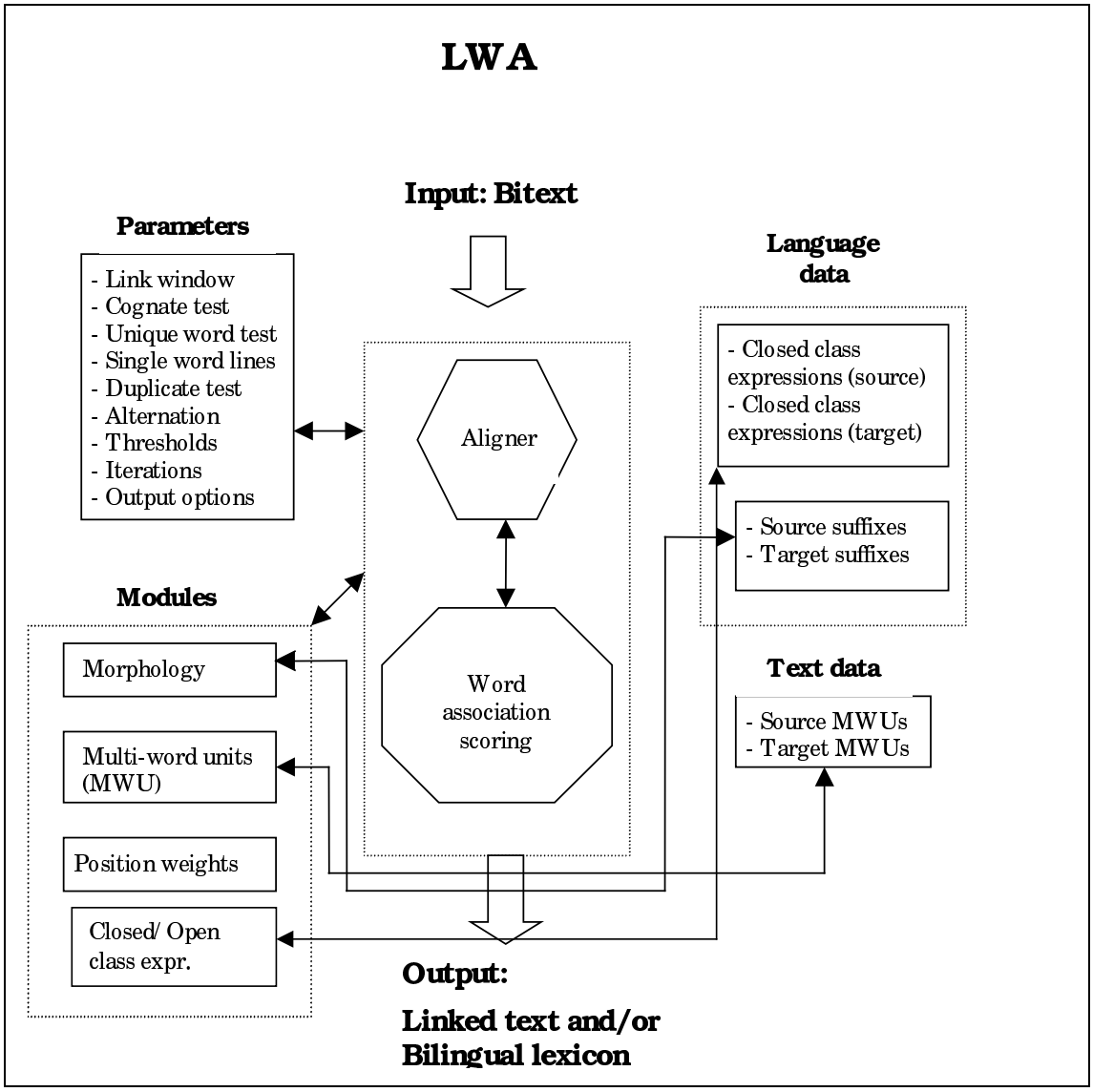


Figure 1. Overview of LWA

The four main modules are:

A *morphological module* that groups expressions that are identical modulo defined suffix sets. Suffix sets are simple lists of suffices that can appear on the same stem as in the examples below:

NOUN1:: {EMPTY, -s}	# book, books
NOUN2a:: {-o, -oes}	# tomato, tomatoes
NOUN2b:: {-ss, -sses}	# mass, masses

A *multi-word unit (MWU) module* that includes multi-word units generated in a pre-processing stage as candidate expressions for alignment.

A *position weight module* that affects the likelihood of a candidate translation according to its position in the sentence; when the weight module is used, the position weights will affect the scores and the ranking.

The *closed/open class expression module* that makes the system observe the division into closed class and open class expressions and the subcategorisation of closed class expressions.

In addition to the above main modules, there are a number of global parameters that can be specified:

Link window. The use of a link window will limit the search in the target segment. If a link window is used and its value is set to 5, this means that five words to the left and five words to the right of the starting position in the target segment will be tried as target candidates.

Cognate test. If the cognate test is used, the user can choose between the Longest Common Subsequence Ratio (LCSR) test (Hunt and Szymanski (1977) and Melamed (1995)) or that the compared units start with *n* identical letters (cf. Simard et al. (1992)). The cognate test is then applied both in the Unique word test (see below) and as a heuristic when choosing among several candidates that show values above the word association threshold.

Unique word test. When the unique word test is used, the bitext will be scanned for unlinked unique tokens (i.e., with frequency 1), and if there are unique candidates on both sides of a bitext segment, these words are then linked.

Duplicate test. When the duplicate test is used, the bitext is scanned for duplicate sentence pairs, i.e., for recurrent sentence pairs where the source sentence and target sentence are identical. In translations with high recurrence degrees, this means that identical tokens of sentence pairs (source and target sentences) will be treated as a single instance.

Alternation. If the alternation parameter is set to true, the linking process will be reversed at the end of each iteration, before the next iteration starts. In other words, when all possible links have been made from source to target, the system tries to find as many links as possible from the target to the source text. If alternation is used together with the morphology module, the possibility to link low frequency source expressions belonging to the same suffix paradigm is increased.

Frequency threshold. This parameter specifies the lowest frequency used in the word association calculation.

Word association score threshold. This parameter specifies the lowest threshold used for the word association score as well as what word association score to be used. The probabilities are estimated by means of the frequency counts.

Number of iterations. This parameter specifies the number of runs the linking process is executed.

The evaluation did not consider all possible combination of options. From previous experience it was known that the highest recall is obtained by using all available modules with little or only slight negative effects for precision.

2.2 The Uppsala Word Aligner (UWA)

The Uppsala Word Aligner is a word alignment system that applies the modular Uplug system (Tiedemann, forthcoming). It is based on earlier studies on bilingual lexicon extraction (Tiedemann 1997, 1998). Special focus was set on the modularity of the system. Hence, the system comprises a set of modules that can be combined and adjusted for specific applications.

The Uplug system is a convenient environment for the combination of single-task modules in order to investigate a variety of different configurations. An application in the Uplug environment is defined by a sequence of modules. A module is any script or program that carries out a specific sub-task. Furthermore, the Uplug system provides an I/O interface (UplugIO) for the manipulation of data from different sources and in different formats. The graphical user interface of the system can be used for the investigation of results in each stage and for the adjustment of parameters and configurations.

UWA is currently the primary application of the Uplug system. Its operation can be divided into three main parts:

1. pre-processing
2. identification and collection of candidate pairs of translation correspondences
3. alignment of instances of translation correspondences

Additionally, a final post-processing stage can be added which includes automatic filters for the exclusion of (obviously) wrong alignments.

Pre-processing. The pre-processing phase accounts for a sub-sequence segmentation of the bitext into link units. A link unit may be a single word unit (SWU) or a multi-word unit (MWU). Pre-processing includes tokenization, the recognition of multi-word units, and the segmentation of the text into link units. Tokenization comprises the separation of tokens from punctuation marks and special characters. This task is not trivial, especially if several languages are to be accounted for. The recognition of multi-word unit can be automated; the process is divided into two sub-tasks, the generation of MWUs, and the identification of their instances in the text. MWUs can be produced via statistical investigations based on frequency counts (cf. Smadja 1993). UWA uses mutual information as association score. Further, constraints concerning the occurrence of function words are applied.

Collection of translation candidates. In this part the system compiles and collects translation equivalents. Several sources and techniques can be used in this collection. In the current implementation, UWA applies the following sources:

- machine readable bilingual dictionaries (MRBD)
- cognate lists (applying string similarity measures)
- pairs of associated word units (applying co-occurrence measures)
- single word bitext segments
- previously aligned word pairs (iteration)

MRDBs from any origin can be used but certainly their quality is decisive for the quality of the word alignment later on. This includes that the chosen MRDBs should be suitable to the type of the text under considerations.

String similarity can be measured by different metrics (Melamed 1995, Borin 1998). UWA uses the Longest Common Subsequence Ratio (LCSR) for this task. Further investigations on the improvement of this metric have been carried out (Tiedemann 1999) but they have not yet been applied in the word alignment process.

UWA supports the same word association scores as LWA (Dice coefficient, mutual information, t-score). The current investigations were focused on the application of the Dice coefficient. Furthermore, simple stemming functions were used in order to reduce the inflectional variety of words in different languages and to improve the statistical calculations.

As noted in sub-section 1.1 special difficulties arise with the usage of multi-word units. UWA supports two different approaches to handle this problem:

- pre-segmentation: the bitext is split into valid link units on both sides; each link unit will be considered as one atomic item
- dynamic segmentation: the system generates all possible link unit candidates by iterative size extension and combines appropriate pairs for the statistical investigations

These techniques can be combined as well.

Word alignment. The actual word alignment is based on the previously collected alignment candidates. Each bitext segment runs through a sequence of single steps. Word alignment candidates can be compiled by associating link units that were identified in the text segmentation process. The alignment is designed to start with the most reliable candidates. Each aligned token is removed from the text such that only non-aligned tokens remain for the next step. In the current stage of the system, 8 alignment steps are defined:

1. align one token units
2. align identical numerics and punctuations
3. align highly similar tokens (lower case)
4. align highly co-occurring tokens (stem form)
5. align pairs from the basic dictionary (lower case)
6. align similar tokens (lower case)
7. align co-occurring tokens (stem form)
8. align remaining one token units

Each alignment step can be adjusted by several parameters. The alignment candidates are ranked by their probability (if an appropriate value is defined, e.g. Dice scores) and the most reliable pairs will be aligned first. Position weights (in form of score reductions) can be used to modify probabilistic scores. Further restrictions can be made in order to reduce the set of alignment candidates. Empirical investigations on the optimisation of parameters settings were made as part of the evaluation.

The Uplug system supports iterative processing. A sequence of modules may run through a number of iterations in order to produce additional results. This technique was extensively used in UWA. As mentioned above, previously aligned word pairs that have been removed from the text can be used to extend the collection of valid alignment candidates. The iteration process can be described as follows:

1. compile a bilingual lexicon from previously aligned words
2. compute alignment candidates by means of word association scores using the remaining tokens in the corpus
3. start the word alignment process all over again including an additional alignment step that applies the newly compiled lexicons
4. continue with (1) until no new alignments can be found

Automatic Evaluation. UWA stores information about each aligned pair. Each aligned unit is represented by a unique identifier corresponding to its origin in the PLUG XML file and its byte span within the text relative to the beginning of the sentence alignment structure. A gold standard was defined for each bitext under consideration. UWA includes an evaluation module that can be used to compare results from a word alignment process with the gold standard. The module produces a protocol with information about each pair from the gold standard and summarizes the alignment result by counting the number of correct, partially correct, incorrect, and not aligned pairs. Finally, evaluation metrics are calculated using those values. Furthermore, information about the actual alignment step is stored for each aligned pair. In this way, the alignment process can be retraced and the quality of each step be investigated.

3 Evaluations

In this section we report the results of the evaluation of LWA and UWA using the method and measures described in sub-sections 1.1 and 1.2.

3.1 Evaluation results (LWA)

The basic setup of the evaluations of LWA was to compare LWA output with the reference data using a separate scoring module that is part of the PLUG Link Annotator. LWA was configured so that all modules, including the cognate and the unique word tests, were used. The frequency threshold was set to 2 and the weighted t-score threshold to 2.5. Each configuration was run in 8 iterations.

All the tests were run on a Compaq Deskpro PC (Pentium III, 500 Mhz, 384 MB RAM) on the Windows NT 4.0 platform.

The results from LWA are reported in table 2 below.

Table 2. Summary of results from running LWA on parts of the PLUG corpus. Rows indicate size of the gold standard (# *golden*), partially correct links (*P*), correct links (*C*), incorrect links (*I*), and missed link units (*M*)

	ensvtxl	ensvtacc	sventscan	svenpeu	ensvfbell	ensvfgord
# golden	500	500	500	500	500[3]	500
C	267	272	203	220	265	228
P	105	109	195	86	75	84
I	62	61	25	75	31	56
M	66	58	77	119	127	132
Recall	0.868	0.884	0.846	0.762	0.744	0.736
Precision I	0.857	0.861	0.941	0.803	0.916	0.848
Precision II	0.736	0.738	0.710	0.690	0.815	0.733
F-measure	0.797	0.804	0.772	0.724	0.778	0.734
Linked types	5980	6770	24208	8996	8639	8353

As indicated in the table recall is high for all the texts (73.6% – 88.4%) whereas precision varies between 80.3 and 94.1% when partial links are considered as correct (Precision I) and between 69 and 81.5% when partial links are valued as 50% correct (Precision II).

Also, the number of retrieved link types can be seen to be quite high. When testing different setups of LWA, it was found that configurations where all the modules and tests were used, increased the number of type links (i.e., the size of the extracted lexicon) by more than 300% compared to when only the statistical core was used. In Table 3 below this is illustrated by comparing the size of extracted lexicons made by the baseline configuration (BASE) and by the ALL configuration on two of the sub-corpora.

Table 3. Size of extracted lexicons for different configurations.

	Size of extracted lexicon (extracted type links)								
Text	**BASE**	**SS**	**WS**	**FS**	**ALT**	**SING**	**PS**	**ALL**	**ALL-NOT-WS**
ensvt-acc	2,179	2,042	2,605	3,663	2,845	4,524	2,428	6,770	6,390
ensvf-bell	2,445	2,152	3,935	4,679	2,727	4,153	2,459	8,639	7,070

The fact that the number of link types increases drastically when all the modules are invoked does not stand out clearly when configurations are compared to a randomly generated gold standard. For example, the automatically calculated recall score for ensvtacc was 81.6% (BASE) and 88.4% (ALL). The differences are not in the actual links made by the system but by the way they are measured. High type recall usually means that a system is better at linking low-frequency items, but in order to capture the

[3] Because two instances in the reference data turned out to be duplicated, the scoring was actually made on 498 instances.

characterisitics of a certain system, it is necessary to vary the strategies for creating samples, or to complement evaluations using randomized gold standards with other methods.

All the gold standards used in the tests were created without restrictions on frequency or categories, except the sventscan text, where function words have been excluded. Two of the texts in the PLUG Corpus (ensvfgord and ensvfbell) were also tested against different gold standards, which had been created with different sampling methods. One type of gold standard was made with a frequency-balanced approach (100 entries with frequency 1-2, 100 with frequency 3-4, 100 with frequency 5-9, 100 with frequency 10-40 and 100 with frequency above 40). The other type of gold standard was also frequency-balanced but contained only content words as input word for the annotation. The results from comparing the system output to these different types of gold standards are illustrative:

Table 4. Recall and precision for the ALL configuration as evaluated by three different gold standards

	ensvtacc		ensvfbell	
Gold standard type	**Recall**	**Precision II**	**Recall**	**Precision II**
A. Random text tokens	0.884	0.738	0.744	0.815
B. Frequency-balanced	0.772	0.736	0.690	0.856
C. Only content words + frequency balanced	0.742	0.768	0.640	0.871

As can be expected, the selection of content words made recall decrease and precision increase. Recall and precision for the ALL configuration when they were evaluated against the three gold standards are shown in Table 4 for (a) random text tokens, (b) frequency balanced words and (c) only content words.

Note that in spite of the large differences the recall and precision data in Table 4 are taken from a single execution of LWA for each text. This means that the sampling strategy used when reference data is created, has a (surprisingly) great effect on the figures for recall and precision. Thus, a system's results when compared to a gold standard containing links that have been collected according to some given criteria, must not be generalised beyond that class of links.

Processing times for LWA varied from 110 minutes (for ensvtacc) to 230 minutes for the largest sub-corpus (sventscan).

3.2 Evaluation results (UWA)

UWA has been applied to both Swedish/English and Swedish/German sub-corpora, although Swedish/English has been covered in greater detail. In this section results for both language pairs are presented.

3.2.1 Swedish/English Word Alignment Results

Each text was processed with a set of different UWA configurations. Each alignment attempt was evaluated by comparing the proposed links with the corresponding reference links in the gold standard. Table 5 below presents the results of the best alignment attempts on each sub-corpus.

Table 5. Summary of alignment results from applying UWA on parts of the PLUG corpus, size of the gold standard (# *golden*), partially correct links (*P*), correct links (*C*), incorrect links (*I*), and not aligned words (*M*)

	ensvfbell	ensvtacc	svenpeu	svenprf	sventscan
# golden	500	500	500	100	499[4]
C	224	232	191	41	222
P	71	87	58	20	150
I	29	45	75	8	41
M	176	136	175	31	86
Recall	62,86	68,07	62,86	68,68	82,55
Precision I	91,04	87,63	76,61	88,40	90,07
Precision II	80,09	75,68	67,69	73,91	71,91
F-measure	70,44	71,67	65,09	71,20	76,86

There are quite remarkable differences between UWA alignments of different texts. The best result over-all was achieved when applied to the Scania corpus (sventscan). Here, the word alignment process produced a much higher recall compared to the other attempts, which is decisive for the high over-all performance on this text. However, the highest precision can be found for the literary text (ensvfbell). Especially the low number of incorrect alignments produces the best performance in this category. The differences in recall are rather small with the exception of the result for the Scania corpus. The lowest precision was measured for the political texts from the European Union. Here, the most freely translated sections can be found. The number of 'null-links' (about 10%) in the gold standard is one measurable reflection of this fact. Another reason for the poor performance might be related to the unspecified translation history of these texts. It is neither known which part of the bitext should be considered to be the origin nor if there was another intermediate language involved in the translation process.

The biggest difference can be found in the number of partially correct alignments compared to the number of identical links and the number of incorrect alignments. The distance between the two precision measures shows different characteristics for each text. In general, alignments from technical texts seem to include a larger number of partially correct alignments compared to political and literary texts (the RF corpus is an exception – the text is very short and its gold standard small). This is certainly due to the larger number of multi-word compounds in technical texts in English compared to the single word correspondences in Swedish.

[4] Due to incompatibilities between the gold standard for the sventscan corpus and the scoring module of UWA one link was lost in the evaluation procedure.

Comparing UWA Configurations

UWA was applied with different parameter settings. Table 6 presents results of UWA alignments with seven different configurations on two bitexts from the PLUG corpus. Here, the base-line configuration (**base**) was extended by additional components such as machine-readable dictionaries (**MRD**), the string similarity module (**sim**), and stemming functions for both languages (**stem**). The last three configurations represent alignment attempts where all UWA modules were included. One of them includes automatic filters (**filter**); another one does not (**no filter**). In contrast to all other UWA alignments the last attempt (**dynamic**) applied dynamic text segmentation instead of pre-segmented texts.

Table 6. UWA alignments using different configurations

	ensvfbell				*sventscan*			
	Precision I	Precision II	Recall	F	Precision I	Precision II	Recall	F
base	89,06	80,85	48,52	60,65	86,62	65,55	68,56	67,02
base+MRD	89,96	80,62	55,48	65,73	87,17	66,66	69,98	68,28
base+sim	88,43	77,72	56,54	65,46	86,18	65,19	72,21	68,52
base+stem	89,7	79,77	51,89	62,88	88,85	67,13	71,6	69,29
all/filter	90,22	79,47	59,28	67,91	89,05	66,71	65,44	66,07
all/no filter	91,04	80,09	62,86	70,44	90,34	68,76	74,44	71,49
dynamic	88,95	77,76	65,4	71,05	90,07	71,91	82,55	76,86

The table above shows clearly the effect of additional modules compared to the base line alignment. Each module produces an improvement in the total performance (considering the F-measure). The alignment attempt including automatic filtering on the Scania corpus represents the only exception in this pattern. The increase in performance is due to major improvements in terms of recall. The table above shows the gain that could be achieved by each extension of the system. However, the differences in precision are more or less insignificant. This fact proofs the quality of the newly discovered links. However, the effect of automatic filtering shows exceptional behaviour. Although a decreasing recall value could be expected, the drop in precision is quite surprising. The filters, which were applied here, seem to exclude a large number of correct alignments, which is very unsatisfying and makes them useless.

Finally, the alignment process including dynamic text segmentation deserves a closer look. The total performances in terms of F-measures represent the best result of all alignment attempts for both text collections[5]. A remarkable gain could be yielded for the Scania corpus by dynamic text segmentation at the expense of computation time. The alignment process took about 16 hours and 20 minutes whereas the results for pre-segmented texts could be achieved after about 4 hours and 15 minutes. However, the difference in computation time of these two approaches is much less significant for smaller texts like the Bellow corpus. The processing time increased here by about 50 % for the UWA alignment when dynamic segmentation was applied. Further investigations will be made in the future in order to improve this approach, which seems to be worthwhile especially for smaller and medium-size text corpora.

[5] The score for precision went down for the Bellow corpus and therefore the alignment with pre-segmentation was selected (measured in terms of weighted F-measures) to represent the best alignment achieved for the Bellow corpus.

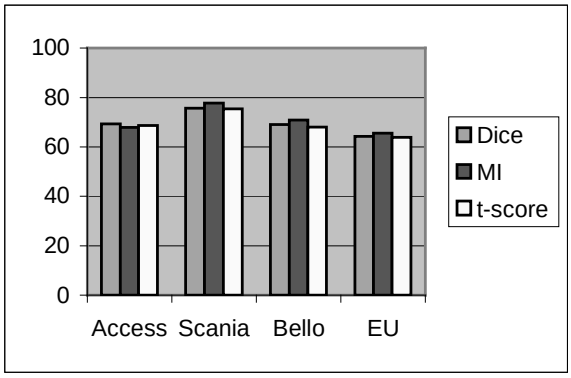
Comparing Different Co-occurrence Metrics

A basic experiment was carried out in order to investigate differences in the alignment results when different association measures are applied. Table 7 shows results, which could be yielded for three different text corpora. Here, very low thresholds were chosen in order to increase recall. In particular, the minimal threshold for the Dice coefficient was set to 0.4, for mutual information scores to 7, and for t-score measures to 1.7.

Table 7. Comparison of three statistical metrics

	sventacc		*sventscan*		*svenfbell*		*svenpeu*	
	Precision II	Recall	Precision II	Recall	Precision II	Recall	Precision II	Recall
Dice	75,28	64,1	72,29	79,31	78,28	61,94	67,71	61,32
MI	71,26	64,73	67,69	91,11	70,58	71,06	58,68	74,46
t-score	63,76	74,33	64,71	90,52	68,36	67,49	54,61	77,03

As can be seen in the table above, the behaviour of the three metrics varies when applied to different texts. A consistent difference between the Dice coefficient and the t-score, however, is that the latter gives higher recall and lower precision. On the other hand, the balance between recall and precision depends on the chosen threshold for the particular metric so general conclusions are hard to draw. The following diagram presents F-measures in order to point out the differences between the alignment attempts:



As can be seen in the picture above none of the three metrics can generally be considered to represent the best metric for the task of word alignment. However, mutual information scores produce the best results for three of the four text corpora. Alignment results based on the Dice coefficient are best for the Access corpus and second best for the others. In general, all results for each of the three scores are very close to each other and they all seem to be applicable to word alignment. However, certain differences can be measured and further investigations have to be made in order to classify characteristics of each statistical metric with regard to word alignment.

Word Alignment and the Extracted Lexicon

Similarly to the LWA evaluations the number of retrieved link types has been investigated for UWA alignments in different configurations. Table 9 shows the resulting counts for alignment attempts on three corpora representing different text types. The attempts include a base-line approach, alignments with dynamic segmentation, and a low-threshold configuration. The number of linked types has been compared with the number of tokens that have been aligned.

Table 8. Type/token ratios for different text types and different alignment configurations

	ensvfbell			*svenpeu*			*sventscan*		
	base-line	dynamic	low	base-line	dynamic	low	base-line	dynamic	low
types	4545	10884	9882	4745	8888	7807	12741	22245	19996
tokens	35306	42378	43081	39528	39342	49295	114309	136673	136251
Ratio	7,77	3,89	4,36	8,33	4,43	6,31	8,97	6,14	6,81

As can be seen in the table above, major increases in the number of linked types could be observed when all modules were included compared to the base-line approach. However, this fact is not similarly reflected in the recall values that were calculated by means of the corresponding gold standard. For example, the recall values for the Bellow corpus were estimated to 48,52% (base-line) and 61,94% (low) whereas the number of linked types in the extracted lexicon was more than doubled. This fact shows the disadvantages of freely sampled gold standards. Furthermore, variations in the type/token ratio could be observed for alignments on texts of different type. The increase of linked types for technical texts is by far not as high as for literary and political texts.

3.2.2 Swedish/German Word Alignment Results

Due to time constraints the Swedish/German parts of the PLUG corpus could not be investigated as thoroughly as their Swedish/English counterparts. However, an overview of alignment results achieved so far is presented in Table 9.

Table 9. Evaluation of Swedish/German alignments

	svdepeu	svdeprf	svdetscan
# golden	500	100	500
C	197	35	233
P	52	12	52
I	85	13	22
M	166	40	192
Recall	59,9	59,18	59,15
Precision I	74,55	78,33	92,83
Precision II	66,76	68,33	84,37
F-measure	63,14	63,43	69,54

The performances of Swedish/German alignments are much lower compared to the alignments of Swedish/English texts. The only exception could be observed in the precision of the Swedish/German alignments from the Scania corpus. However, the recall value here is much lower and therefore the total performance is lower as well compared to the Swedish/English counterpart. These results are quite surprising; the relation between Swedish and German is usually considered to be very close. Especially the similarities in the usage of compositional compounds imply a potential raise in precision as well as in recall. One reason for the drop in performance can be found in the complex morphology of the German language, even though a stemming function was used for German as well. However, the quality of the German stemming module (which is freely available as a Perl extension) was not examined in particular. Although the total performance is lower even in precision, it can be seen that the gap between

Precision I and Precision II is much smaller than for Swedish/English texts. This is qmost probably due to the more similar usage of compounds in both languages. In this way a remarkable improvement could be observed for alignments of the technical text (Scania). Closer investigations have to be carried out in the future in order to evaluate differences between alignments of texts from both language pairs in order to compare the performances more systematically.

4 Analysis of results

Looking at the numbers, which were presented in Tables 2 and 5 of the previous section, we can first see some consistency in the results on different sub-corpora. Both systems achieve their best results (in terms of F-values) on technical texts. This is especially due to a much higher recall on those texts. The worst result for both systems was produced for the texts from the European Union. This can be explained by their complexity: the translation is more free and there is a considerable number of null links in the reference data (10%).

In general, LWA tends to yield higher recall values than UWA. This can possibly be explained to a large extent by LWA using the t-score and UWA the Dice coefficient (with the set thresholds). As was illustrated in Table 7, the t-score gave a consistently higher recall than the Dice coefficient with a threshold set to 1.7, while the weighted t-score threshold for LWA, 2.5, is comparable to a non-weighted threshold of 1.65.

As for precision, the difference between the two systems is negligible, with UWA coming on top on LWA for two sub-corpora and LWA on top of UWA for the other two. Thus, the decline in precision that affected UWA when using the t-score (again, according to Table 7) seems not to affect LWA to the same extent. A reason for this might be LWA's special treatment of closed class words (see below for further details.)

It can be seen that the performance differences are not the same for all texts. Although the results for the Scania corpus are very similar in almost every evaluation category the values for e.g. the EU texts differ remarkably especially in recall. In order to investigate this phenomenon it was decided to inspect the actual links that were produced for each item from these two gold standards. In this way it was expected to find some patterns and explanations for those differences that were observed. This was done by collecting the alignments proposals of both systems and the annotations from the gold standard in a single table. Then the proposed alignments were compared in terms of correspondences and variations between the two systems and the gold standard.

First, the most similar result is considered: alignments from the Scania corpus. The following numbers have been counted for the combination of alignment types (**C**-correct, **P**-partially correct, **I**-incorrect, **M**-misalignments):

Table 10. LWA and UWA alignment types from the sventscan corpus

sventscan		*UWA*			
		C	*P*	*I*	*M*
LWA	C	142	24	6	28
	P	41	105	16	30
	I	7	5	5	4
	M	25	16	9	25

Both systems produced quite similar results for this text collection. The numbers above reflect this fact in terms of the number of alignment type combinations. There are not many variations between both systems. The biggest differences can be found in pairs for which one system produces correct alignments and the other system only partially correct alignments, and in pairs where one system misses one link totally whereas the other system produces at least a partially correct alignment.

Now, more detailed investigations were made in order to analyse these differences but also the results both systems have in common. Both systems produce twice the same incorrect alignment (out of five). Furthermore, about 40% of partially correct alignments are identical for both systems. The most common part that is missing in partial alignments is the definite article in English expressions (about 14% of the partially correct alignments for LWA and about 25% for UWA). However, the two systems often miss different instances of the definite article. Only 12 instances of a missing article are common for partially correct alignments proposed by the two systems. Table 11 shows the parts that were missed at least twice in partially correct aligned units by one of the systems.

Table 11. Missing parts in UWA and LWA alignments from the sventscan corpus

LWA		UWA	
Source	Target	Source	Target
2 [* av]	2 [* valve]	3 [* av]	7 [The *]
2 [* bort]	6 [The *]	3 [* i]	6 [a *]
3 [* i]	2 [a *]		3 [are *]
2 [Sätt *]	5 [are *]		4 [is *]
2 [hastighet*]	2 [be *]		31 [the *]
2 [kan*]	2 [by *]		
	5 [is *]		
	2 [may*]		
	22 [the *]		

As can be seen in the table above both systems suffer from definite noun constructions in English and from the usage of auxiliary and particle verbs.

Another reason for partiality is that the systems generate phrases that are longer than those defined in the gold standard. For brevity, this will be referred to as inclusion in the sequel. In the following table the number of such alignments is shown:

Table 12. The number of too long alignment pairs from the Scania corpus

sventscan	*Source*	*Target*
LWA	4	10
UWA	24	30

Here though, a clear difference between UWA and LWA can be observed. The numbers of inclusions in UWA alignments are much higher than for LWA. UWA tends to split the text into parts larger than the smallest unit that can be linked. These alignments have to be considered to be partially incorrect even though correct alignments can be found among them. In this way, 9 of the 24 alignment pairs, for which the LWA proposal was identical with the reference link in the gold standard but UWA proposal was marked as partially correct, could be observed to be correct regarding to the text segmentation that was done by the system. The result of the automatic evaluation highly depends on the characteristics of the gold standard. The gold standard that was used here defines the minimal translation units that could be found instead of phrasal constructions. However, the inclusion of such phrasal structures can be very interesting for several purposes such as machine translation (Sågvall Hein, forthcoming). The alignment approach that was used by UWA here included iterative (dynamic) phrase generation. In this way, a large number of candidates is considered which does not rely on pre-compiled mono-lingual collocations. This method seems to support the recognition of multi-word units though it tends to generate phrasal structures that are longer than those found in a gold standard based on minimal relations of equivalence.

Another type of partial alignment is when all parts of a phrase are properly linked to each other but the phrase as such was not recognized. Here, evaluation metrics with regard to partiality can be used for score approximations.

The descriptions above consider the alignment results for the attempts where UWA and LWA could achieve the most similar performance. In order to compare both systems in the case of outcome differences the attempts with the biggest divergence, the alignment approaches on EU texts, were investigated. Similarly to the evaluation table for the Scania alignments UWA and LWA results were merged together with the gold standard.

Table 13 shows result type combinations.

Table 13. LWA and UWA alignment types from the EU corpus

svenpeu		*UWA*			
		C	*P*	*I*	*M*
LWA	C	135	2	28	44
	P	6	44	10	35
	I	9	7	21	23
	M	11	5	16	73

As mentioned earlier, the biggest difference in the total performance of different alignment attempts can be found in the recall value. The table above illustrates this fact quite well. There are quite a large number of at least partially correct alignment proposals from the LWA system in cases where no link was found by the UWA approach. Furthermore, a rather large number of incorrect proposals by UWA system were linked correctly by the LWA. However, the EU corpus and its gold standard have

some special characteristics, which have to be considered when evaluating alignment proposals. First, there are much longer sentences included when compared to technical texts. In this way, instances of similar word types are more often included in the same sentence. Investigations on alignment proposals for units, which were linked correctly by LWA but where UWA proposed a wrong alignment, showed that about 30% of those mistakes are due to the choice of the wrong instance of the correct word type. Another characteristic of the EU corpus is the large number of so-called 'null-links', items that do not have corresponding parts in the translation. Further investigations on the same set of alignment proposals showed that another 35% of the links, which were proposed by UWA, represent links for source language items that have been marked as 'null-linked' items in the gold standard, in 90% of the cases functional words.

It can be stated that LWA produced a large number of additional links that were missed by UWA. Linköpings Word Aligner seems to try on a larger set of candidates and even though a number of incorrect alignments were added a large number of at least partially correct links could be gained. As argued above, the use of the t-score with a rather low threshold may be responsible for this fact. Among the correct LWA proposals that were missed by UWA a number of correct aligned function words can be found. This represents about a third of those alignments. A possible explanation for this ability might be the classified function word lists that are used in the LWA system.

Similarly to the Scania corpus there are several types of partially correct alignments. Table 14 shows the number of inclusions that could be counted in the alignment results of the EU texts.

Table 14. The number of too long alignment pairs from the EU corpus

svenpeu	*Source*	*Target*
LWA	8	9
UWA	2	12

Compared to the Scania results the values above show a very different behaviour of UWA with regards to the EU corpus. This is probably due to the segmentation method that was applied here. As opposed to the Scania alignment pre-compiled collocations were used in order to split the EU texts into link units. This technique seems to reduce the number of inclusions. However, this text type is very different from the technical text in the Scania corpus with all its repetitions of multi-word units. Further investigations have to be made in order to study this phenomenon.

Finally, we will take a close look at partially correct alignments. Table 15 summarizes the most common parts (frequency > 1) that were missed in partially correct proposals by each alignments system.

Table 15. Missing parts in UWA and LWA alignments from the EU corpus

LWA				UWA			
Source		Target		Source		Target	
2	[* att]	5	[The *]	3	[* att]	2	[and*]
8	[av*]	5	[be *]	2	[bli*]	2	[be *]
4	[de *]	3	[is *]	2	[de *]	3	[for*]
2	[de*]	2	[of the *]	2	[enligt*]	2	[no*]
2	[för*]	17	[of*]	2	[för*]	2	[of the *]
2	[informations- och *]	29	[the *]	2	[inom*]	2	[of*]
2	[inom*]	2	[will*]	3	[kan*]	2	[proposed*]
2	[kommer att*]			2	[och*]	2	[such as*]
3	[om*]					20	[the *]
2	[på*]					2	[will*]

As can be seen in the table above a larger number of Swedish particles were missed compared to the counts on partial links from the Scania corpus. This is due to the different characteristics of political texts compared to technical texts. However, similarly to the Scania alignments the most common part that is missing in English alignment units is the definite article '*the*'.

In summary, both systems have similar difficulties and generally show a similar behaviour when applied to different text types. However, we have found that UWA tends to over-generate inclusions when dynamic text segmentation is applied. This might be useful for several applications such as example-based machine translation. However, dynamic text segmentation is slow because of the large amount of candidates. LWA has a consistently higher recall for each text it was applied to, something which may be due to its use of the t-score for this evaluation. However, it tends to include many partial alignments in early stages of the alignment process, which makes it harder to improve the precision. The LWA system is fast and robust and can be easily adapted to new language pairs. UWA system applies the Uplug system, which provides graphical user interfaces and a set of convenient tools for experimentation, configuration, and investigations of intermediate and final results.

5 Discussion

5.1 Evaluation measures

First, some improvements on evaluation techniques for word alignment systems based on the experiences during our investigations shall be mentioned. As stated earlier, automatic evaluation of word alignment results is not a trivial task. Difficulties arise especially with partially correct alignments. Within the evaluations described above simple approximations of recall and precision were applied. However, these metrics are rather unsatisfying with regard to partiality. Another approximation was defined for the word alignment competition that was initiated by ARCADE. In this competition precision and recall were defined as follows:

C_{trg} – number of overlapping tokens in system proposal and gold standard
S_{trg} – number of target tokens proposed by the system
G_{trg} – number of target tokens in the gold standard

$$\forall G_{trg} > 0 : Q_{gold} = \frac{C_{trg}}{G_{trg}}, \forall G_{trg} = 0 : Q_{gold} = 0$$

$$recall_{ARCADE} = \frac{\sum Q_{gold}}{n(I) + n(P) + n(C) + n(M)}$$

$$\forall S_{trg} > 0 : Q_{system} = \frac{C_{trg}}{S_{trg}}, \forall S_{trg} = 0 : Q_{system} = 0$$

$$precision_{ARCADE} = \frac{\sum Q_{system}}{n(I) + n(P) + n(C) + n(M)}$$

However, these definitions do not seem to be suitable for our purposes. First, the evaluations are based on the proposed translation only. However, the recognition of correct source language link units is one of the tasks of the systems that were described here. Using the measure proposed by ARCADE a partially correct unit on the source language side would be counted as completely correct. Secondly, the definition of precision that was used in the ARCADE competition seems to be unfair. Here, even pairs that were missed by the systems are included. In this way, the precision value also depends on the quantity of the results proposed by the system although precision should describe the quality of obtained results only. Furthermore, the ARCADE definition of precision does not take care of inclusions. Although inclusions will modify the recall value, they will be counted as completely correct in the calculation of precision. This does not seem to fit to the task of word alignment. For this reason we propose another metric for the approximation of precision and recall for word alignment systems:

C_{src} – number of overlapping source tokens in (partially) correct link proposals, C_{src}=0 for incorrect link proposals
C_{trg} – number of overlapping target tokens in (partially) correct link proposals, C_{trg}=0 for incorrect link proposals
S_{src} – number of source tokens proposed by the system
S_{trg} – number of target tokens proposed by the system
G_{src} – number of source tokens in the gold standard
G_{trg} – number of target tokens in the gold standard

$$Q_{partial} = \frac{C_{src} + C_{trg}}{\max(S_{src}, G_{src}) + \max(S_{trg}, G_{trg})}$$

$$recall_{align} = \frac{\sum Q_{partial}}{n(I) + n(P) + n(C) + n(M)}$$

$$precision_{align} = \frac{\sum Q_{partial}}{n(I) + n(P) + n(C)}$$

Using the definitions above, partially correct alignments are considered in both measures proportionally to the number of words that describe the difference between the gold standard and the proposed alignment. Inclusions are similarly included in the

precision value as well as links that miss a part compared to the gold standard. Consider the examples in table 16 for a better understanding.

Table 16. Precision and recall for partial links - example

	Source	target	$Q_{partial}$
gold standard	Reläventil TC	TC relay valve	
proposed	Reläventil TC	Relay valve TC	3/5 = 0.6 2/5 = 0.4
gold standard	ordinarie	ordinary	
proposed	ordinarie skruv	ordinary bolts	2/4 = 0.5
gold standard	kommer att indikeras	will be indicated	
proposed	kommer att indikeras	will the indicated	2/6 = 0.33 0/6 = 0 2/6 = 0.33
gold standard	vill	wants	
proposed	-	-	0
gold standard	Scanias chassier	Scania chassis	
proposed	Scanias chassier	chassis Scania	2/4 = 0.5 2/4 = 0.5
		precision recall	(3.17)/4 = 0.79 (3.17)/5 = 0.63

The examples in the table above demonstrate the behaviour of the newly defined evaluation measures with regard to some special link types that may occur in word alignment results. They illustrate clearly the ability of these measures to handle partially correct proposals in cases of inclusions as well as in cases of missing parts. A final problem remains: Alignments that were proposed in terms of sub-links may be twisted as shown in the last example in the table above. The measures do not take care of this phenomenon but consider them to be completely correct. However, $precsion_{align}$ and $recall_{align}$ can be considered to be the most precise approximations for the purpose of alignment evaluations when partially correct links are included.

5.2 Design issues

Both LWA and UWA are designed to be modular and include a large variety of knowledge-lite parameters. Many improvements were made and the systems were further developed as a result of the co-operation in the PLUG project. Apart from many similarities, there are also some differences in the approaches taken by the two systems.

The UWA approach comprises three independent stages. In each stage a set of data is produced which can be used in the next stage. As opposed to the LWA system, alignment candidates are collected in bilingual lexicons before the actual alignment starts. Various approaches to automatic lexicon extraction are applied for the collection of alignment candidates. This collection of lexicons is then used for the linking process in the alignment phase. Therefore, quality and size of each lexicon determines the performance of the word alignment later on. However, candidates are ranked and controlled by various parameters in order to improve the quality of the alignment. Due

to the general architecture of UWA, the alignment is independent of the direction of the translation. The bitext is considered to be a set of two strings without any directed relation.

LWA clusters words in groups of similar frequency. The alignment process starts with high frequent words and iterates down to low frequent words. Furthermore, functional words and lexical words are considered in different steps. In this way, words from those classes can never be mixed in proposed links. The morphological module is used within the alignment stage. Candidate pairs, which were ranked on the top, have to run through a simple suffix test, in order to group word-forms together and to improve the co-occurrence statistics. Due to this principle the alignment direction is important. Results will be different when the bitext is swapped. Therefore, a change of alignment direction can be included in LWA alignments. Furthermore, similarity tests can be added in order to find additional links. An important parameter is the position weight, which modifies statistical scores.

Both systems use pre-compiled collocations (based on co-occurrence investigations). However, the application is slightly different. While LWA uses both the collocations and the single words that are included for its investigations, UWA does not allow breaking up valid collocations into smaller pieces. An exception is the approach to dynamic segmentation. Here, phrases will be compiled within the extraction process instead of monolingual pre-segmentation. Furthermore, both methods can be combined as well.

LWA and UWA comprise sets of scripts and modules, which were mainly written in Perl. LWA was tested on Sun Solaris and Microsoft Windows whereas UWA was implemented on a Linux platform. UWA also requires Perl/Tk in order to run the graphical user interface. Furthermore, the system supports standard UNIX database managers like GNU DBM or SDBM and provides possibilities to connect to relational database managers via the transparent DBI module. The LWA parameters are set in specific configuration files, which control the complete alignment process. The system itself can be started in the command line mode. UWA comprises a set of configuration files, which are used to set parameters for each module that is included. Parameter settings as well as intermediate results can be inspected from the graphical user interface. The alignment system itself can be started from the same interface and runs as a background process as specified.

The implementation of the LWA system was focused on the task of word alignment. The system represents a complex toolbox comprising several modules that are specialized on this task. It is fast and robust and simple to adjust to new language pairs. However, the system is command line oriented and the possibilities for the user to interact with the system are minimal. It requires an experienced user to adjust parameter settings for additional experiments. A user with less experience is probably happier with a simpler version of LWA that have fewer options, but which makes use of all modules with default settings, so that a consistently high recall is ensured.

UWA system is based on the Uplug toolbox, which was developed for the work on text corpora. The system is focused on the development of reusable modules not necessarily for the task of word alignment. Special attention was paid to the management of data collections in different formats. In this way, UWA modules can be applied to textual

data from various sources and for different purposes. Furthermore, the graphical user interface is a convenient platform for the adjustment of parameter settings for specific applications. Intermediate results of each module can be examined directly from the interface and conversions into supported data formats are provided.

The common word alignment system to be developed should combine the advantages of both systems. We propose to merge the two systems and their modules and provide a common user interface. The interface shall be based on the GUI of the Uplug system and both systems will be integrated on a common platform. The amount of adjustable parameters will be reduced to the set of decisive parameters that can be modified by the user. The system will be developed to run on Linux and Sun Solaris systems and possibly on Microsoft Windows as well. The implementation will be mainly based on Perl implementations and some common Perl extensions. The system requirements in particular will be specified in the user manual.

6 Conclusions

The evaluations in the sections above describe differences and similarities between the two systems. Suggestions for several improvements can be concluded for both systems from these investigations. As a result, a common system as a combination of both will be developed.

The evaluation experience reported in this paper has shown, we believe, that LWA and UWA both are useful word alignment systems that can be used as a basis for the creation of a freely available research system. While partly built on similar philosophies and sub-processes, they also complement one another in other respects. Thus, by providing both of them under the same graphical interface in simplified versions that do not sacrifice performance too much, a very useful tool can be provided to various research communities interested in parallel text processing.

As for the evaluation process itself we have found certain short-comings in our original proposal. We still find that gold standards are an important resource for the evaluation of word alignment systems, but they must be designed so as to meet the different needs of different kinds of systems. We also conclude that the measures used in this evaluation were not optimal, but have also made a proposal for new improved measures of recall and precision.

References

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

Borin, L, 1998. Linguistics isn't always the answer: Word comparison in computational linguistics. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen.

Fung, P. and Church, K. W., 1994. K-vec: A New Approach for Aligning Parallel Texts. In *Proceedings from the 15th International Conference on Computational Linguistics* (Coling-94). Kyoto: 1096-1102.

Melamed, I. D., 1995 Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*, Boston/Massachusetts.

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.

Merkel, M. and Ahrenberg, L., 1999. Evaluating word alignment systems. PLUG Report.

Merkel, M., Andersson, M., and Ahrenberg, L, forthcoming. The PLUG Link Annotator - Interactive Construction of Data from Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Merkel, M., 1999. Annotation Style Guide for the PLUG Link Annotater. Linköping. PLUG report, Linköping University.

Sågvall Hein, A., forthcoming. The PLUG Project: Parallel corpora in Linköping, Uppsala, Göteborg: Aims and achievements. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University.

Simard, M., Foster, G. F., and Isabelle, P., 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal: 67-82.

Smadja, F., 1993. Retrieving Collocations from Text: XTRACT. *Computational Linguistics 19(1)*.

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

Tiedemann, J. 1998. Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen.

Tiedemann, J., 1999. Automatic Construction of Weighted String Similarity Measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Coropora*, University of Maryland, College Park/MD, 1999.

Tiedemann, J., forthcoming. Uplug - A Modular Corpus Tool for Parallel Corpora. In *Proceedings from the Parallel Corpus Symposium*, April 22-23, 1999, Uppsala University, Sweden.

Véronis, J. and Langlais, P., forthcoming. Evaluation of parallel text alignment system - The ARCADE project. To be published in *Parallel Text Processing*. J. Véronis. Berlin, Kluwer.