

# The PLUG Link Annotator – Interactive Construction of Data from Parallel Corpora

Magnus Merkel, Mikael Andersson & Lars Ahrenberg

Department of Computer and Information Science  
Linköping University  
S-58183 Linköping

## Abstract

*In this paper an approach of using gold standards to evaluate word alignment systems is described. To make the process of creating gold standards easier, an interactive tool called the PLUG Link Annotator is presented along with the Link Scorer, which automatically evaluates the output from a word alignment system against the gold standard. It is argued that using reference data in this manner has several advantages, the most important being consistency in evaluation criteria as well as savings in time, due to the fact that the reference data only need to be constructed once, but can be applied many times.*

## 1. Introduction

One of the most valuable uses of a parallel corpus is for the generation of bilingual concordances. Current interactive tools make correct alignment of parallel texts at the sentence level a fairly quick process and the aligned text can then be searched for a word or word pattern to generate a concordance. Some systems can even be consulted over Internet e.g. the Gothenburg Pedant Corpus (Ridings 1998) and the RALI TransSearch system (Simard et al. 1993).<sup>1</sup>

These systems will high-light the given word for you, but are not yet able to locate the word or words that it corresponds to on the other side. Word alignment is a more difficult problem than sentence alignment and the current systems are not able to perform at a level near that for sentence aligners. The best system in the recent ARCADE word alignment contest had a precision of 77% (Véronis and Langlais 1999), while several sentence aligners had a precision of over 95%. These figures were obtained by comparing system performance with a Gold Standard, that is a set of reference data compiled by human annotators.

To evaluate the output from a word alignment system, reference data can be constructed before the actual alignment takes place as a kind of *prior reference*. Such reference data are sometimes referred to as *gold standards* and are usually a sample of the bitext that has been prelinked manually by one or several annotators and then used to test the alignment output automatically. *Posterior reference* on the other hand is when the output from a system is given to annotators who, following specific instructions, evaluate the output and annotate the whole output or a sample thereof for correctness and completeness.

Using posterior reference does not entail the creation of tailor-made software. It is sufficient that a sample of the system output is evaluated after the alignment. However, as each reference data has to be created every time the system has been run, the evaluation will have to start from scratch each time the system has been used.

---

<sup>1</sup> The Pedant system can be found at <http://eupc30.adm.gu.se/~ridings/pedant.html> and TransSearch at <http://www-rali.iro.umontreal.ca/TransSearch/TS-simple-uen.cgi> (February 1999).

An existing bilingual lexicon can also be used as prior reference for testing the performance of bilingual lexicon extraction. The disadvantage of using lexicons as a gold standard is that there may be problems in coverage; a standard bilingual dictionary will, for example, not contain domain-specific terminology. Furthermore, as bilingual lexicons commonly only list the base form of words, the output from the alignment system must be lemmatised.

Setting up a gold standard before the system is used, is definitely more efficient due to the savings in time. One gold standard can be used to check hundreds of sets of output data from one or several systems automatically. Another advantage concerns consistency; as the system output will be evaluated against the same reference data, the risks of having inconsistent evaluations will be minimised. Of course, there can be manual mistakes in the reference data, but at least the same flaws will be present in all applications of the reference data. The drawback is that annotation guidelines as well as software for the annotation of the gold standard and the scoring have to be created, but once this is done, the advantages will outweigh the disadvantages.

The production of gold standards can also be helped by using interactive tools. The PLUG<sup>2</sup> Link Annotator is such an interactive tool that has recently been developed. In this paper we will present the system and the considerations that underlie its design.<sup>3</sup> Primarily, it is developed with the aim of evaluating the word alignment programs used in the PLUG project, (Ahrenberg, Andersson and Merkel 1998) and Tiedemann (1998), but it is not tailor-made for these systems and could therefore also be used to evaluate other word alignment systems. Moreover, with minor modifications the PLUG Link Annotator could be adapted for annotation of other correspondence characteristics within translations studies and contrastive linguistics. In addition, we will discuss the annotation guidelines that we favour for use with the system.

## 2. Related work

Annotated bilingual data were recently used in two different projects, the Blinker project and the ARCADE project, with the same overall purpose, namely to acquire a more objective way of evaluating word alignments. In the Blinker project (Melamed 1998) a dedicated visual tool was developed that makes the annotation of the parallel Bible texts simple. The annotator connects the different tokens in the text by drawing lines on the screen. With the Blinker tool bilingual annotation is performed on all the tokens in a sample of sentence pairs. Tokens could be linked to "null word" on the other side, but annotators were forced to make a choice for each token and could not indicate uncertainty. In the ARCADE project (Véronis & Langlais 1999) annotation was made in a bilingual document by selecting the correspondences in the text. A selection of single word tokens is taken as the starting point for the annotation. The annotator could also give a confidence level (graded on a scale from 0 to 3) and indicate the correspondence type (*normal*, *omission*, *referring expression*, *spelling error*, etc.).

The PLUG Link Annotator approach resembles the ARCADE way of annotating bilingual data, in the way that both approaches use a sample of source words from the bitext. The difference between the ARCADE and PLUG approaches is that in the first application of the PLUG annotation, all the input words are sampled randomly from the source text, whereas in the ARCADE project the source words were selected from a certain frequency range and chosen for their polysemic properties. However, the basic principles remain the same.

---

<sup>2</sup> PLUG stands for Parallel Corpora in Linköping, Uppsala and Göteborg, a project jointly funded by Nutek and HSRF under the Swedish National research program in Language Technology.

<sup>3</sup> We are indebted to Anna Sägvall-Hein and Jörg Tiedemann for valuable discussions on the properties of the PLUG Link Annotator.

### 3. The PLUG Link Annotator

The PLUG Link Annotator is a piece of software that is run interactively to create reference word lists, which can be used to measure the performance of a word alignment program automatically. The input to the PLUG Link Annotator consists of a list of source words together with the source sentences where they occurred and the corresponding target sentences. In the current version, we use a random selection of 500 words for each bitext, but the choice of input words could be made differently in the pre-processing stage. For example, one could decide to pick out words from a certain frequency range, ignore function words, or select words from certain specified categories, if parts-of-speech information is available.

The architecture of the PLUG Link Annotator facilitates extensions of this kind. Figure 1 below illustrates how different criteria can be used for selecting the setup of reference data.

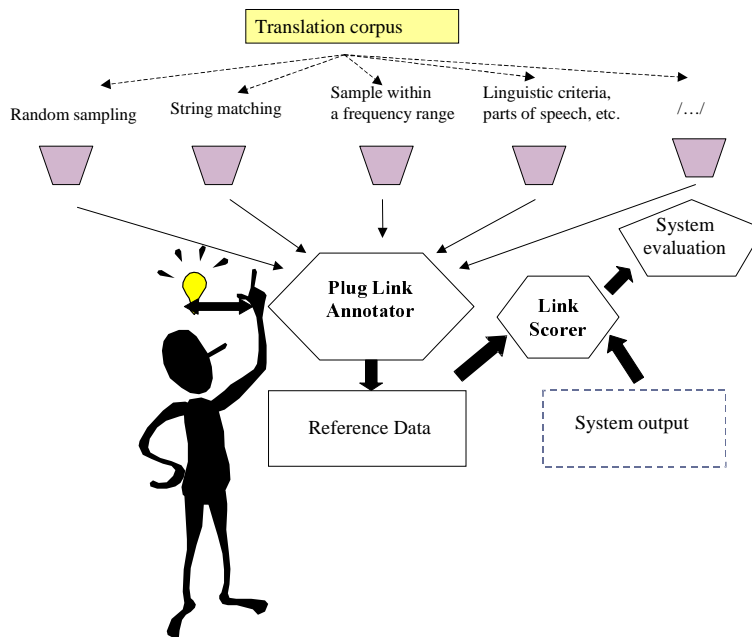


Figure 1. Overview of how the PLUG Link Annotator and the Link Scorer are used for evaluation of word alignment systems.

The reference data is created in the PLUG Link Annotator by a human annotator. Finally the Link Scorer compares the output from the word alignment system with the reference data and returns evaluation data.

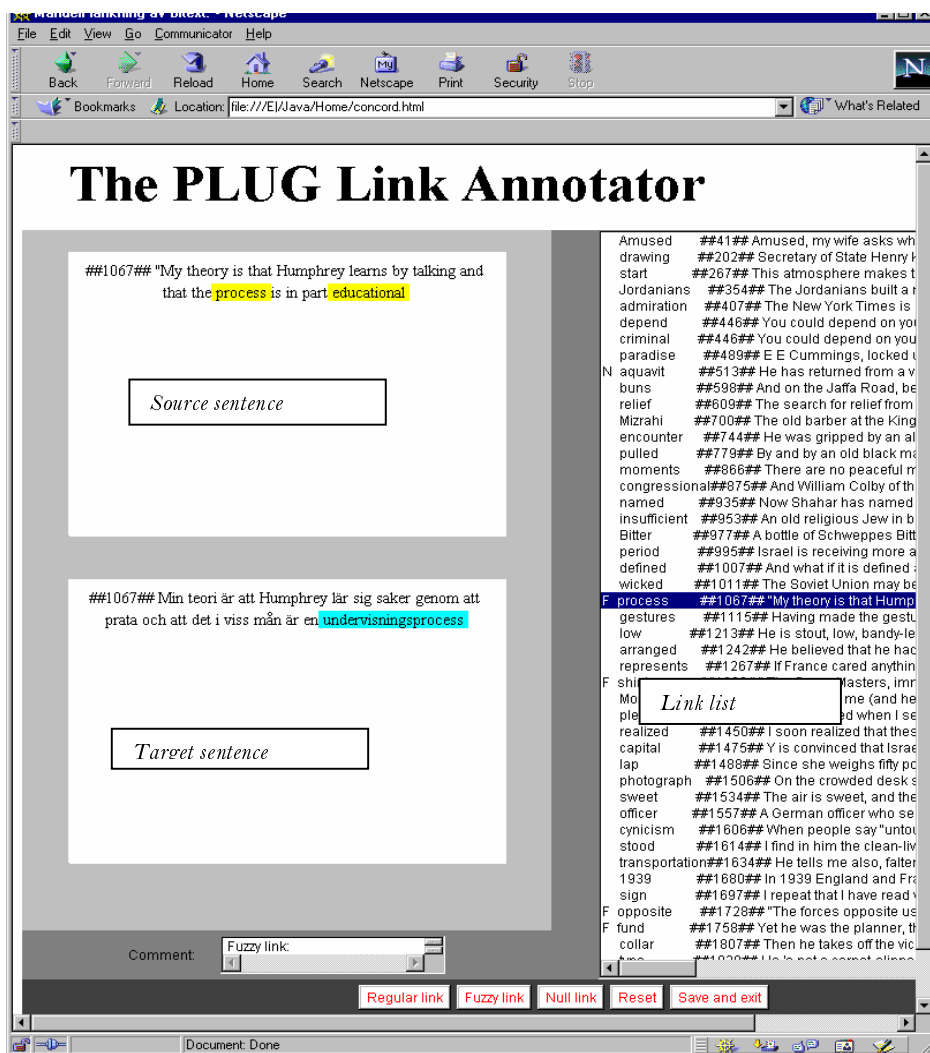
While the purpose of a word alignment program is often to generate lexical data, the principal annotation task can be considered to be *textual linking*, which means that the goal is to find correspondences between tokens present in the source and target text. It is important to stress that the objects of interest here are the translations and correspondences as they are manifested in the actual texts. Lexical links on the other hand can be seen as derivatives of textual links, after the application of some filters: function words are excluded and only base forms of words are listed.

The PLUG Link Annotator is accessed through a web browser, as is illustrated in Figure 1.

The interface consists of four major fields:

1. The source sentence field in the upper left corner (where the original source word to be annotated is highlighted).
2. The target sentence field (where the target candidates are to be selected by the user).
3. An action bar at the bottom consisting of buttons for different commands.
4. A scrollable list of links that have been created in the session so far.

Every time a source word is presented, the user has to choose at least one option in the action bar. If the correspondence is straightforward, the user selects the corresponding target word(s) and clicks on “Regular link”. If there is no translation of the target word, the user selects “Null link”. The selection is done by clicking on the left mouse button. If the user wants to deselect an item, this is done by clicking on the selected item again. If the relationship between the source word and the target word is regarded as “fuzzy”, the user has to indicate this by clicking “Fuzzy link”.



Action bar

Figure 2. The PLUG Link Annotator interface

#### 4. Guidelines for the annotators

In order to acquire consistent annotations when several annotators are involved it is necessary to create a document where general and specific guidelines for the annotation work are set up. The guidelines used for the evaluation of LWA are presented in Merkel (1999). The starting point is a single word on the source side, and the task is to select the best two-way correspondence starting from this word. Two general guiding rules were adopted from Véronis (1998):

1. Mark as many words as necessary on both the target and source side.
2. Mark as few words as possible on both the target and source side.

To ensure that there is a two-way equivalence, as many words as necessary should be selected. Even if the starting point is always the source word, the selected parts in the two texts should correspond in both ways.

Below the notational convention of indicating the sampled starting word in **face** is used. The possible extension of the source word and the preferred target word(s) are shown in **bold face**.

SOURCE: For more information on configuring a particular **SQL** database server, search Help for “ODBC drivers”...

TARGET: Mer information om hur du konfigurerar en viss **SQL-databasserver** finns i Hjälp under “ODBC-drivrutiner”...

Given that the initial source word is “SQL”, it is straightforward to see that there is no single word that corresponds to the source word in the target sentence. The target expression to be selected must be “SQL-databasserver” which means that the source unit has to be extended to “SQL database server”.

The annotator must also decide whether a particular link is *regular* or *fuzzy*. The main division between fuzzy and regular links has to do with meaning. If the source and target units are different in degrees of specification or are semantically overlapping in some sense, the link should be considered as fuzzy. For example, the units “came out - “hade vågat sig ut” do not carry exactly the same meaning in “The spiders came out from behind their pictures” and “Spindlarna hade vågat sig ut från sina tillhåll”. Therefore the link “came out” - “hade vågat sig ut” is a fuzzy link.

There are two additional principles that determine whether a link is categorised as *fuzzy* or *regular*:

**Inflectional principle:** Change of inflectional form (but with the same parts-of speech) is considered to be a *regular link*. For example, changes in number, tense, definiteness and voice are considered to be regular links.

**Categorial principle:** Change of parts-of-speech (e.g. from verb to noun) is considered to be a *fuzzy link*. For example, in the pair “A snort of pleasure” - “En förtjust nysning”, the prepositional modifier “of pleasure” corresponds to the Swedish adjective “förtjust” as a fuzzy link.

Other more specific guidelines concern the annotation of omissions (null links), phrasal expressions, verb constructions and infinitive markers, pronouns, proper names, terms, articles, noun phrases, etc.

In some cases there will not be any target word(s) that correspond to the source word. When a source unit does not have an equivalent textual unit on the target side, this is indicated by using the button “Null link”.

The standard strategy for handling omissions can therefore be expressed as follows:

**OMISSION RULE 1:** *If a source word or phrase does not have a textual counterpart on the target side (either partial or whole), the link should be classified as a “null link”, i.e., an omission.*

An example can illustrate Omission Rule 1:

SOURCE: **Setup** installs the ODBC files and ODBC icon in Control Panel.

TARGET: ODBC-filerna installeras och ODBC-ikonen placeras i Kontrollpanelen.

In the example above the word “Setup” is not translated in the target sentence (the voice is changed to the passive and the agent is deleted).

It must be stressed that although the general principles for making annotations are similar for different language pairs, the specific guidelines for handling for example verb constructions and noun phrases may differ across language pairs. Consequently, specific annotation guidelines for each language pair of interest have to be constructed.

## 5. The Link Scorer

The output of the PLUG Link Annotator is a text file that consists of information needed to automatically calculate measures of the quality of the output from a word alignment program. For each entry, there is information on what sentence pair the entry belongs to, the initial source word and its character position, the type of units (single word or multi-word), the type of link (standard, fuzzy or null), etc. An example of an entry in the gold standard file is shown below. Here the initial word that the user has been asked to link is *traffic*, which resulted in the source unit that was selected becoming *network traffic* and the corresponding target unit *nätverkstrafiken*.

```
align ID:    224
sample:     129|7
word:       traffic
link:       network traffic -> nätverkstrafiken
link type:  regular
unit type:  multi -> single
source:     121|7 & 129|7
target:     134|16
source text:##224## To do that, you add a system
table named MSysConf to the SQL database and make
entries in the table that control network traffic.
target text:##224## För att kunna göra optimeringen
lägger du till en systemtabell med namnet MSysConf i
SQL-databasen och för in värden som styr
nätverkstrafiken.
```

Figure 3. Entry in the output from PLUG Link Annotator

When a word alignment system’s output is checked against the gold standard (the PLUG Link Annotator file), precision and recall figures are calculated automatically. The dedicated program for doing the scoring is called the *Link Scorer*. By scoring the results in this manner, it is possible to compare the performances of different systems. With data from the scoring

phase, it is possible to pinpoint both strong and weak points of the systems, for example how the systems perform on multi-word units and fuzzy links.

Another important use of the Link Scorer is to optimise the configuration of a word alignment system internally. If some of the gold standards developed with the PLUG Link Annotator, are used as training data, it would be possible to experiment with different configurations and parameters of a system, in order to find the optimal combination of, for example, search order, function word lists, collocation data, statistical thresholds and co-occurrence scores.

An example of the output from the Link Scorer is shown in Table 1.

Table 1. Output from the Link Scorer

<b>Number golden:</b>	500 (Regular: 388, Fuzzy: 26, Null: 86)
<b>Number identical:</b>	272 (R: 207, F: 2, N: 63)
<b>Number partially linked:</b>	109 (R: 100, F: 9, N:0 )
<b>Number completely different:</b>	61 (R 29, F: 9, N: 23)
<b>Total number tried:</b>	442
<b>Number not tried:</b>	58 (R: 52, F: 6, N: )
<b>Recall:</b>	0.884
<b>Precision I</b>	0.862
<b>Precision II</b>	0.739
<b>F-measure:</b>	0.805

The number of links in the golden standards is given (500) as well as information on the number of regular, fuzzy and null links. The tested system has found 272 identical links, 109 partially correct links (with some overlap), and 61 system links were found to be wrong compared to the gold standard. Recall is here given as 88.4 per cent (number tried/number of golden links). Two kinds of global precision scores are also given:

$$Precision I = \frac{\text{occur}(\text{identical links}) + \text{occur}(\text{partial links})}{\sum \text{occur}(\text{identical links, partial links, different links})}$$

$$Precision II = \frac{\text{occur}(\text{identical links}) + (0.5 \times \text{occur}(\text{partial links}))}{\sum \text{occur}(\text{identical links, partial links, different links})}$$

In the first precision measure partial links are considered to be correct, and in the second partial links are scored as 0.5, that is half of an identical link.

A value for F-measure is also given, that is, the harmonic mean of recall and precision:

$$F - \text{measure} = 2 \frac{\text{precisionII} \times \text{recall}}{\text{precisionII} + \text{recall}}$$

The above measures and data are only examples of what the Link Scorer can present. For example, if translation spotting is to be evaluated, it is possible to calculate the ARCADE variants of recall and precision. More detailed information could also be obtained by using scores that are related to the qualitative differences between regular and fuzzy links.

## 6. Examples

In this section two applications of using the PLUG Link Annotator are presented. The first example shows how an evaluation of different system configurations of word alignment program can be performed. The second application illustrates the effects of using different criteria for selecting the input words for the PLUG Link Annotator

### 6.1 Evaluation of different system configurations

To be able to evaluate the Linköping Word Aligner (Ahrenberg, Andersson & Merkel 1998, Merkel 1999b) in a more objective way it was decided to use the PLUG Link Annotator and the Link Scorer. In this section, two of the texts in the PLUG corpus are evaluated, namely *Microsoft Access User's Guide* and Saul Bellow's *To Jerusalem and Back*. The first stage in the evaluation contained the following steps:

1. Create Gold Standards for the different translations in the Linköping translation corpus<sup>4</sup> with the aid of the PLUG Link Annotator.
2. Run the Linköping Word Aligner (LWA) system with a number of configurations of different modules.
3. Evaluate the different configurations with the Link Scorer.
4. Based on step 3, use the best module configuration and test what the effects are of changes to numerical parameters, such as frequency thresholds, initial frequency, t-values, position weights and size of the link window.

The first step involved sampling out 500 source words for each text and then using the PLUG Link Annotator to create the gold standards. The sampling was done randomly from the source texts (token sampling). Two annotators annotated the texts independently based on the guidelines described earlier (Merkel 1999a). The inter-rater agreement using the PLUG Link Annotator was between 89.8 and 95.2 per cent (counting all annotations) for the four texts, which indicate the annotator guidelines had been used and worked for the purpose. The inter-rater agreement was calculated as the proportion of exactly identical links from two different annotators in relation to the total number of links in the reference data.

The second stage of the evaluation consisted of running the LWA system on the two different texts with different configurations of modules. The following configurations were tested:

1. Baseline, only t-scores (BASE)
2. Baseline and function word subcategorisation (SS)
3. Baseline and position weights (WS)
4. Baseline and morphology (FS)
5. Baseline and reverse linking direction (alternation) (ALT)
6. Baseline, single word lines test and unique word test (SING)
7. Baseline and multi-word units (PS)
8. All modules and all tests (function word categorisation, position weights module, morphology module alternation, single word lines test, unique word test and cognate test) (ALL)
9. All modules and tests except the position weights module (ALL-NOT-WS)

The first configuration only used the core statistical machinery while in the configurations 2-7 different modules were added to the statistical machinery. In configuration 8 all modules and tests were used, and in configuration 9 everything except the module that uses information

---

<sup>4</sup> The whole novel by Bellow was linked in this test, but the Access translation was shortened to approximately the same size as the Bellow translation in order to have comparable bitext sizes.

about the relative positions of words in the sentences was applied. All the configurations used the same global parameters, for example:

- Frequency threshold: 2
- T-value threshold: 1.65 without weights and 2.5 with position weights.
- Number of iterations: 8

Furthermore, LWA used the same resource files (suffix lists, subcategorised function words, and MWU lists) for the different configurations (when applicable).

The measures used for calculating the precision were of two types: (1) *precision I*, that is, partial links are considered correct, (2) *precision II* meaning that a partial link is considered as “half-correct” (contributes with 0.5 for the precision score), see also section 5.

When the Link Scorer produced the results from all the configurations, it was obvious that the scores seemed to be very similar. Consider the data for recall, precision and F-measure in Table 2:

Table 2. Recall, precision and F-measure for nine configurations of LWA (using 500 random text tokens).

Text		BASE	SS	WS	FS	ALT	SING	PS	ALL	ALL-NOT-WS
Access	Recall	0.816	0.804	0.808	0.862	0.896	0.826	0.828	<b>0.884</b>	0.872
	Prec. I	0.838	<b>0.898</b>	0.824	0.865	0.814	0.837	0.833	0.861	0.871
	Prec. II	0.726	<b>0.784</b>	0.710	0.744	0.691	0.726	0.719	0.738	0.751
	F-measure	0.768	0.794	0.756	0.799	0.780	0.773	0.770	0.804	<b>0.807</b>
Bellow	Recall	0.630	0.590	0.692	0.688	0.672	0.668	0.654	<b>0.744</b>	0.738
	Prec. I	0.920	<b>0.955</b>	0.901	0.892	0.898	0.912	0.913	0.916	0.928
	Prec. II	0.842	<b>0.877</b>	0.815	0.810	0.817	0.824	0.836	0.815	0.828
	F-measure	0.721	0.706	0.748	0.744	0.737	0.738	0.734	<b>0.778</b>	0.755

The assumption was that the highest scores would turn up in the two rightmost columns (ALL and ALL-NOT-WS). The reason behind this assumption was that the addition of different modules and tests would improve the performance when they were added to the simplest baseline configuration. The recall figures above support this assumption; the ALL configuration has the highest recall (0.884 for the Access text and 0.744 for the Bellow text). However SS (baseline plus subcategorised closed-class words) contain the highest precision for both translations. The best scores for F-measure are found in the ALL configuration (Bellow) and in the ALL-NOT-WS configuration (Access).

The scores seemed slightly mysterious at first, but when looking closer at them and also evaluating the size of the bilingual lexicons produced, it was clear that the data in Table 2 may not represent the performance of the different system configurations accurately. For example, recall for Access using the alternation configuration (ALT) is 0.896 whereas the recall using the ALL configuration is 0.884, that is the ALT configuration and the ALL configuration are almost identical. If the number of generated type links (bilingual lexicon entries) for these configurations is compared, the difference is, however, clearer. The ALT configuration produces a bilingual lexicon with 2,845 entries whereas the ALL configuration creates 6,770 lexicon pairs. The same pattern appears for both texts, see Table 3 below.

Table 3. Size of extracted lexicons for each configuration

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

The reason for the differences in recall has to do with the random selection of text tokens in the gold standard. The sampling of random text tokens results in a preference for high frequency word types in the reference data. Consequently, if it is “easier” to link high frequency units accurately with a less sophisticated machinery, then most configurations will score well on the randomly selected text tokens.

## 6.2 Evaluation of LWA using different types of gold standards

To investigate if the problem of seemingly “small” differences between simple configurations and more complex ones was connected to the random text tokens present in the gold standard, the selection of source words which were fed into the PLUG Link Annotator was redone in a different manner. This time a frequency-oriented approach was used where the sampled source items were divided into five groups of different frequency (f): (1) f=1-2, (2) f=3-4, (3) f=5-9, (4) f=10-40, and (5) f>40, where each group holds 100 source tokens, in total 500 samples. The assumption here was that the preference for more or less only picking out high-frequency words would be avoided and thereby better represent the performance of the different system configurations. In addition it would provide a handle for observing the capacity of LWA in different frequency ranges. Table 4 below summarises the scores for recall, precision and F-measure for the total 500 source tokens in the gold standard.

Table 4. Recall, precision and F-measure for nine configurations (using frequency-balanced text tokens).

Text		BASE	SS	WS	FS	ALT	SING	PS	ALL	ALL-NOT-WS
Access	Recall	0.616	0.598	0.682	0.598	0.646	0.672	0.618	<b>0.772</b>	0.74
	Prec. I	0.823	<b>0.856</b>	0.807	0.826	0.823	0.798	0.818	0.842	0.838
	Prec. II	0.727	<b>0.751</b>	0.701	0.725	0.724	0.702	0.728	0.736	0.738
	F-measure	0.667	0.666	0.685	0.709	0.683	0.687	0.668	<b>0.753</b>	0.739
Bellow	Recall	0.500	0.444	0.552	0.57	0.572	0.544	0.502	<b>0.690</b>	0.600
	Prec. I	0.931	<b>0.971</b>	0.921	0.911	0.832	0.921	0.921	0.958	0.952
	Prec. II	0.771	0.820	0.824	0.800	0.703	0.806	0.766	<b>0.856</b>	0.837
	F-measure	0.607	0.576	0.661	0.666	0.603	0.650	0.607	<b>0.764</b>	0.699

As can be seen in the table above, all scores are lower than when the random sampling of text tokens was used (see Table 2), but this is expected since the frequency-balanced gold standard will contain a higher proportion of low frequency tokens which are harder to align. However, using all the modules (ALL) is definitely the best option according to this gold standard; the ALL configuration receives the highest recall and F-measure. Using the subcategorised function words (SS) will actually produce a higher precision, but the SS recall is considerably lower than the ALL recall, which will favour making the ALL configuration as the preferred choice.

By looking closer at the different frequency ranges of the sample words in the gold standard, it is possible to observe where the strength and weaknesses of different modules lie. In Table 5

below, recall, precision and F-measure for the five different frequency ranges are presented for the ALL, BASE and SS configurations.

Table 5. Recall, precision and F-measure for three different configurations (frequency-balanced)

Text		f=1-2	f=3-4	f=5-9	f=10-40	f>40
Access (ALL)	Recall	<b>0.460</b>	<b>0.850</b>	<b>0.78</b>	<b>0.88</b>	<b>0.89</b>
	Prec. I	0.826	<b>0.835</b>	<b>0.910</b>	0.818	0.820
	Prec. II	0.685	<b>0.753</b>	<b>0.782</b>	0.739	0.719
	F-measure	<b>0.673</b>	<b>0.762</b>	<b>0.826</b>	0.766	0.745
Access (BASE)	Recall	0.220	0.600	0.58	0.840	0.840
	Prec. I	0.818	0.800	0.793	<b>0.857</b>	0.845
	Prec. II	0.727	0.717	0.690	0.756	0.744
	F-measure	0.338	0.653	0.630	<b>0.796</b>	<b>0.789</b>
Access (SS)	Recall	0.24	0.58	0.600	0.820	0.750
	Prec. I	<b>0.875</b>	0.828	0.817	0.854	<b>0.907</b>
	Prec. II	<b>0.750</b>	0.741	0.717	<b>0.756</b>	<b>0.793</b>
	F-measure	0.364	0.651	0.653	0.787	0.771
Bellow (ALL)	Recall	<b>0.480</b>	<b>0.580</b>	<b>0.710</b>	<b>0.800</b>	<b>0.880</b>
	Prec. I	0.979	0.983	<b>0.986</b>	0.913	0.932
	Prec. II	<b>0.917</b>	0.836	<b>0.887</b>	0.794	0.847
	F-measure	<b>0.630</b>	<b>0.685</b>	<b>0.789</b>	<b>0.797</b>	<b>0.863</b>
Bellow (BASE)	Recall	0.050	0.270	0.530	0.750	0.900
	Prec. I	<b>1.000</b>	0.926	0.925	0.907	0.900
	Prec. II	0.600	0.815	0.811	0.807	0.822
	F-measure	0.092	0.406	0.641	0.777	0.859
Bellow (SS)	Recall	0.050	0.230	0.480	0.680	0.780
	Prec. I	<b>1.000</b>	<b>1.000</b>	0.979	<b>0.941</b>	<b>0.936</b>
	Prec. II	0.600	<b>0.913</b>	0.885	<b>0.846</b>	<b>0.859</b>
	F-measure	0.092	0.367	0.622	0.754	0.818

The data show that the definite strength for using all the modules is accentuated when low-frequency words are compared. Consider for example the recall figures for the Bellow novel when the ALL configuration has been used compared to BASE and SS. Only five of the 100 tokens with frequency 1 or 2 are linked with the BASE and SS configurations, but the ALL configuration manages to link 48 of the 100 tokens present in the gold standard of this frequency range. The suspicion vented earlier that a simpler machinery (such as BASE and SS) will actually perform relatively well on high frequency tokens is confirmed by the fact that the relative differences between the different systems decreases with higher frequency.

A third kind of gold standard was also developed against the configuration where all modules and the default parameters were used (ALL) was tested. This time only content words were selected as input words to the PLUG Link Annotator. As for the second type of gold standard, the selection of words was divided into the five different frequency ranges. 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 6 for (a) random text tokens, (b) frequency balanced words and (c) only content words:

Table 6. Recall and precision for the ALL configuration as evaluated by three different gold standards

Gold standard type	Access		Bellow	
	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

Consequently, recall and precision will vary depending on the type of gold standard used. Note that the recall and precision data in Table 6 are taken from one execution of LWA. The links and lexicons produced are therefore the same; it is the different strategies for selecting the reference data that are different.

The use of reference data can be complemented by data from the extracted lexicons. Information on how many lexical entries that have been extracted will shed a different light on the recall scores from the automatic scoring. For example, the data given in Table 3 (size of extracted bilingual lexicons) provide the information that type recall has more than tripled in the ALL configuration compared to the BASE configuration.

The evaluation has shown the application of using the PLUG Link Annotator and the Link Scorer to automatically compare different configurations of a word alignment system. The scores for recall and precision can differ significantly depending on what kind of selection strategy is used for the input words to the gold standard. Here it has been shown that a random word sampling of source text tokens will not show the different strengths and weaknesses inherent in different configurations as clearly as a frequency-balanced sampling of input words. The reason for this is that a random text token selection will favour the selection of high frequency words, which in turn are easier to align with a less sophisticated machinery. When more low-frequency words were included the relative differences between different setups of LWA appeared more clearly. To make the characteristics of a word alignment system (or configuration) even clearer, one could design other types of selection criteria, for example, word type based selection or selection based on grammatical criteria.

## 7. Conclusions

In this paper we have addressed a solution to evaluating the performance of word alignment systems with the help of an interactive annotation tool, the PLUG Link Annotator and a program (the Link Scorer) that automatically measures recall and precision for a word alignment system. The use of prior reference data (or gold standards), such as the ones produced by the PLUG Link Annotator have clear advantages as it means that the reference data only have to be produced once, but can be applied several times in order to compare the performance of different word alignment systems or different configurations of the same system.

In practical evaluations it was shown that the criteria for selecting the input words to the PLUG Link Annotator will influence precision and recall scores. By evaluating different system configurations of the Linköping Word Aligner it was possible to show which configuration was the best for several different texts and also the kind of qualitative differences that different modules contributed with (for example in different frequency ranges).

## References

- Ahrenberg, L., M. Andersson & Merkel, M. (1998). *A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts*. Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montréal, Canada, 10-14 August 1998, 29-35.
- Melamed, I. D. (1998). "Annotation Style Guide for the Blinker Project" , IRCS Technical Report #98-06, University of Pennsylvania.
- Merkel M. & L. Ahrenberg (1998) Evaluating Word Alignment Systems. PLUG report, Department of Computer and Information Science, Linköping university.
- Merkel, M. (1999a) "Annotation Style Guide for the PLUG Link Annotator". PLUG report, Department of Computer and Information Science, Linköping university.
- Merkel, M. (1999b). *Understanding and Enhancing Translation by Parallel Text Processing*. Ph.D. Thesis 607. Department of Computer and Information Science, Linköping University, Linköping.
- Ridings, D. (1998). "PEDANT: Parallel Texts in Göteborg." *Lexikos* 8: 243-268.
- Simard, Michel, George F. Foster, and Francois Perrault. (1993) *TransSearch: A Bilingual Concordance Tool*. Centre for Information Technology Innovation, Laval.
- Tiedemann, Jörg. (1998) "Extraction of Translation Equivalents from Parallel Corpora ". In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen, 1998.
- Véronis, J. & P. Langlais (1999). "Evaluation of parallel text alignment systems". In *Parallel Text Processing* (ed. J. Véronis), under publication, Kluwer.
- Véronis, Jean. (1998) "ARCADE - Tagging guidelines for word alignment" . Aix-en-Provence: Univeriteté de Provence. URL: <http://www.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/>.