

# Annotation Style Guide for the PLUG Link Annotator

Version 1.0

March 19, 1999

Magnus Merkel

*Linköping university*

**Abstract:** This report describes the PLUG Word Annotator, a piece of software that makes it possible to create a reference list of corresponding source and target units in a bitext. The input to the program is a list of randomly chosen source words and the contexts in which they occurs. The human annotator's task is to decide the correspondences between units in the source and target and to determine whether the link should be categorized a standard link (default) or as fuzzy link. The main part of the report deals with specific guidelines on how to use the PLUG Link Annotator when the source text is English and the target text is Swedish.

## 1. ABOUT THIS GUIDE

This document was written as an aid for people who use the PLUG Word Annotator, a piece of software that is used 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 Word Annotator consists of a list of source words together with the source sentence where the sampled word occurs and the corresponding target sentence. In the current version, we use a random selection of 500 words for each corpus, but the choice of input words could be made differently in the preprocessing 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.

Below two entries in the input file are shown. The two words that should be annotated are "include" and "objects".

110, 111: include::1  
 ##SOURCE## If you plan to use Microsoft Access with data that's stored in SQL databases on your network, you must include Open Database Connectivity (ODBC) support as one of your Setup options.  
 ##TARGET## Om du tänker använda Microsoft Access tillsammans med data som finns i SQL-databaser i nätverket, måste du vid installationen också ta med alternativet ODBC (Open Database Connectivity).

116, 129: objects::1  
 ##SOURCE## (The database administrator must also use security commands to tell Microsoft Access that you have permission to use the objects.)  
 ##TARGET## Databasadministratören måste också använda säkerhetskommandon och på så vis definiera vilken behörighet du har att använda objekten i databasen.

The numbers before the source word give information on the sentence ID and the character position for the source word in the source sentence. The PLUG Word Annotator is presented in a HTML page which looks as follows:

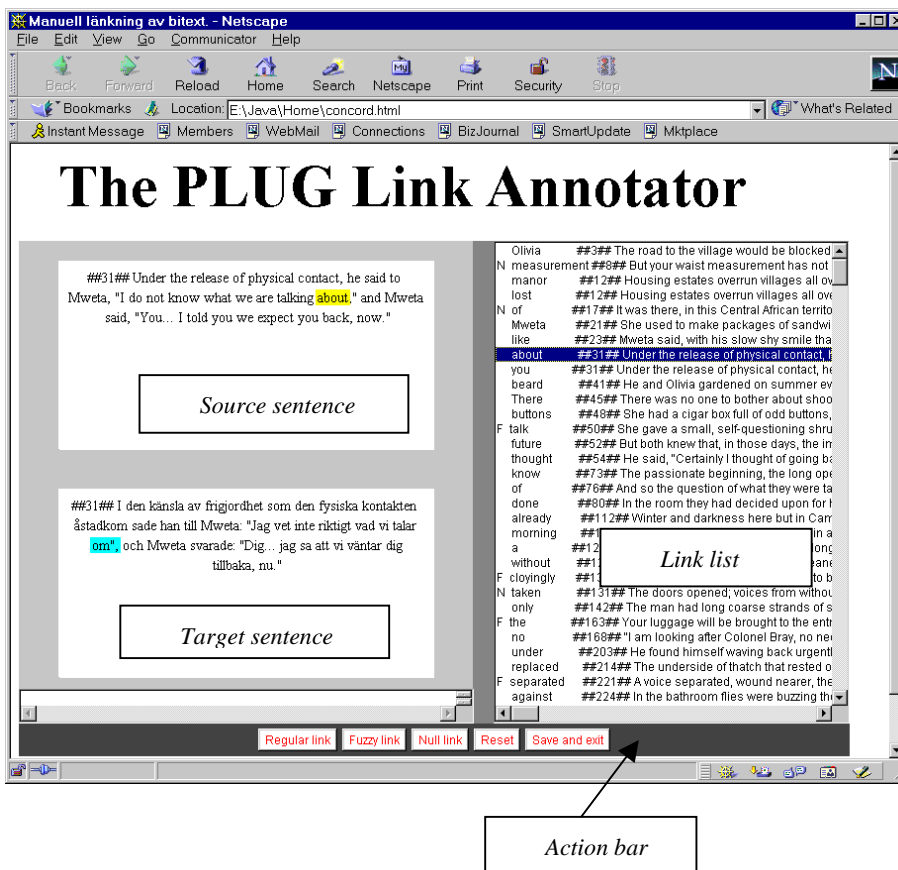


Figure 1 The Plug Link Annotator

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 marked in yellow).
2. The target sentence field (where the target candidates are to be selected by the user). The target words become blue when the user clicks on them.
3. A list of all previously made annotations are shown on the right hand side.
4. An action bar at the bottom consisting of buttons for different commands.

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 a shift in meaning or the link is considered to be non-standard, the “Fuzzy link” should be chosen. If there is no translation of the source word(s), the user selects “Null link”. Clicking on the left mouse button does the selection. If the user wants to deselect an item, this is done by clicking on the selected item again. The original input word can be brought back by clicking on the “Reset” button. It is not necessary to complete the whole annotation in one session; it is possible to save the annotated links that have been made by selecting “Save and exit”.

## 2. OTHER ANNOTATION PROJECTS

The principal annotation task is to identify *textual links*, which means that the goal is to find the correspondences between tokens present in the source and target text. The result of an annotation session is an output file with textual links that can be checked against the output from a word alignment program, thereby testing the system’s performance on textual linking.

*Lexical links* on the other hand can sometimes be seen as a generalization of textual links, where the emphasis lies on creating precise and general bilingual lexical units. If a word alignment program is going to be tested for its performance on lexical linking, the textual linking gold standard can be modified for this task, for example, by removing articles and certain other function words.

In the PLUG Link Annotator the source words presented to the user are randomly chosen from the bitext, which means that the annotation does not involve all words in the sample. In the Blinker project (Melamed 1998) the bilingual annotation is performed on all the tokens in a sample of sentence pairs. The PLUG Link Annotator approach resembles the ARCADE way of annotating bilingual data (Veronis & Langlais 1999), in that single word

tokens are taken as the starting point for the annotation, but it is the annotator's task to decide if the source unit should be extended to comprise more than one word. The difference between the ARCADE and PLUG approaches is that for 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.

### 3. GENERAL GUIDELINES

In textual linking the source token corresponds to zero, one or several words in the target text. The two major rules of thumb when annotating textual links can be summarized as follows (Veronis 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 is necessary should be selected. Even if the starting point always is the source word, the selected parts in the two texts should correspond in both ways.

In the examples we use the notational convention of indicating the sampled starting word in **underlined bold face**. The possible extension of the source word and the preferred target word(s) are shown in **bold face**.

#### **Example 1:**

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":

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"...

**Example 2:**

SOURCE: When you tell Microsoft Access which sale you're interested in, Microsoft Access can look up **the phone number** based on the relationship between the two tables.

TARGET: När du talar om vilken försäljning det rör sig om letas **telefonnumret** upp med hjälp av relationen mellan tabellerna.

If the given source word is “phone” then the corresponding unit on the target side must be “telefonnumret”. As this also contains the definite article we must return to the source side and extend the selection to “the phone number”.

Remember that the goal is to preserve the two-way equivalence, so the rule number two will apply if smaller units can carry the equivalence.

**Example 3:**

SOURCE: In version 1.x, setting a control's **Visible** property to No ...

TARGET: Om du i version 1.x anger egenskapen **Synlig** för en kontroll till Nej ...

Here there is no need to extend “Visible” on the source side as there is a clear correspondence to “Synlig” on the target side.

**Example 4:**

SOURCE: ... to run the program from **a** server ...

TARGET: ... att köra programmet från **en** server ...

In example 4 there is a simple and straightforward equivalence between the indefinite articles in the source and target sentences. However, for English-Swedish the definite article “the” as the target word will often cause an extension of the selection on the source side as the definite article in most cases cannot be separated from the stem in Swedish:

**Example 5:**

SOURCE: ... to run **the program** from a server ...

TARGET: ... att köra **programmet** från en server ...

**Choosing between fuzzy or regular links**

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 which determine whether a link is categorized 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.

## 4. DETAILED GUIDELINES

### 4.1 Omissions in translations

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:

**Example 6:**

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

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

In example 6 the word “Setup” is not translated in the target sentence (the voice is changed to the passive and the agent is deleted). Omissions like the above are annotated as “null links” by clicking on the “Null link” button.

Some other examples of omissions:

**Example 7:**

SOURCE: Can Microsoft Access find all the answers using the information in your tables?

TARGET: Finns svaren i tabellerna?

The English word “all” is not translated and is therefore regarded as a null link.

**Example 8:**

SOURCE: in a few special cases

TARGET: i vissa fall

**Example 9:**

SOURCE: you do n’t start by using the QBE grid

TARGET: börjar du inte med frågerutnätet

Do-constructions are generally regarded as omissions. In example 9 the correspondences are between “start” and “börjar”, and between “n’t” and “inte”.

**Example 10:**

SOURCE: showing the relationship between the two tables

TARGET: som relaterar tabellerna till varandra

In example 10 the whole segment is paraphrased in such a way that the meaning of the source word is no longer explicitly expressed and therefore we regard the link as a “null link”.

The guideline for determining whether a certain link should be regarded as a fuzzy link (in a paraphrase) or as an omission can be stated as follows:

**OMISSION RULE 2:** *If a word or phrase is paraphrased in such a way that the meaning for the word or phrase is not expressed in the target, the link should be classified as a “null link”, i.e., an omission.*

The same rule will apply in example 11–15 below:

**Example 11:**

SOURCE: for details on available formats, see chapter 8

TARGET: en beskrivning över tillgängliga format finns i kapitel 8

**Example 12:**

SOURCE: You 'd rather **avoid** the tedious task of calculating each change

TARGET: I stället för att ägna värdefull tid åt att ändra och uppdatera posterna var för sig

**Example 13:**

SOURCE: To make them look **more** uniform, you may want to size them to the grid.

TARGET: Du kan ge dem ett enhetligt utseende genom att fästa dem mot rutnätet.

**Example 14:**

SOURCE: You 'd have to do everything **their** way, no options given.

TARGET: Det är de som bestämmer hur allting ska göras, det finns inget att välja på.

**Example 15:**

SOURCE: His nose **is** nobly hooked, and slender.

TARGET: Han har en smal och ädel kroknäsa.

## 4.2 Phrasal constructions

Phrasal constructions, for example idioms, fixed expressions and multi-word abbreviations are to be treated as single units on both the source and target side. The standard rule for handling phrasal constructions can be expressed as follows:

**PHRASAL CONSTRUCTION RULE:** *If the given word is a part of standard phrase (idiom, fixed expression etc.), the whole phrasal construction should be selected.*

Some examples:

**Example 16:**

SOURCE: Some controls, **such as** text boxes, combo boxes and check boxes, have an attached label.

TARGET: Vissa kontroller, **te**x textrutor, kombinationsrutor och kryssrutor, har en kopplad etikett.

**Example 17:**

SOURCE: So that, as they say in Chicago, is where the smart money is.

TARGET: Det är alltså där de smarta pengarna bör placeras, **som det heter** i Chicago.

**Example 18:**

SOURCE: This document went **back and forth**.

Detta dokument passerade **fram och tillbaka**.

**Example 19:**

SOURCE The United States is, **after all**, the prime revolutionary country.

TARGET: USA är **när allt kommer omkring** det ursprungliga revolutionslandet.

### 4.3 Proper names and terms

Multi-word units that can be regarded as referring expressions, for example proper names and specific terms should be treated as single units.

***PROPER NAMES RULE 1:** If a word is part of a multi-word referring expressions (name of a person, building, institution, company and product), then the whole multi-word unit should be treated as a single unit. For example, New York Times, Microsoft Word, Empire State Building, United States of America, Henry Kissinger, etc.).*

***PROPER NAMES RULE 2:** If a word is part of a multi-word predicative (or descriptive) expressions or functions as a classifier for another unit (for example a proper name), then the word should be treated as a different unit. For example: “the **Save as** button” consists of the name of the button “Save as” and the classifier (the) “button”.*

Some examples:

**Example 20:**

SOURCE: Microsoft Access kan display values such as numbers ...

TARGET: I **Microsoft Access** kan värden som nummer ...

Here the initial word “Microsoft” is part of the product name “Microsoft Access” and therefore there has to be an extension to cover the whole name.

Consequently, when a proper name has been deleted, the whole proper name should be selected as part of the “null link” as can be shown in example 21 below.

**Example 21:**

SOURCE: To specify the display format **Microsoft Access** uses for data in a field, you set the Format property.

TARGET: Du anger visningsformat med egenskapen Format.  
(Null link)

#### 4.4 Noun phrase constructions and articles

Noun phrases and the use of articles requires special guidelines. For example, the difference between how definiteness is expressed in English and Swedish makes specific rules necessary. As units can be discontinuous the distinctions between different parts of noun phrase constructions can still be maintained. Let us look at some examples:

**Example 22:**

SOURCE: if you want more space to enter or edit **a property setting**, press ...

TARGET: om du vill ha mer utrymme att skriva eller redigera **egenskapen** på, trycker du på ...

In the above example the indefinite article in English is not expressed in the Swedish target. Instead the corresponding Swedish noun phrase is marked as definite. This is however not by itself a reason to classify the link as fuzzy, as changes in number are still regarded as regular links (see Inflectional principle on page 6). The link should be nevertheless be marked as “fuzzy” because there is a difference in specification degree between the target unit and the source unit (the information that it has to do with the “setting” is not present in the translation).

Other examples:

**Example 23:**

SOURCE: the format affects only how **a value** is displayed ...

TARGET: formatet påverkar endast hur **värdet** visas, inte hur det lagras i tabellen (regular link)

**Example 24:**

An asterisk stands for any number of characters in the same position as the asterisk.

Asterisken står för alla tecken i samma position som asterisken.  
(regular link)

***NP/ARTICLE RULE 1:** If the definite article in English has a corresponding independent definite article in Swedish (de/det/de), the article as well as the head noun should be selected if the head noun is marked for definiteness. Adjectival modifiers and quantifiers should not be included in the link.*

Three examples will illustrate NP/Article Rule 1:

**Example 25:**

SOURCE: the underlying field name

TARGET: det underliggande fältnamnet

**Example 26:**

SOURCE: The American role in the war of 1973 has been widely misunderstood.

TARGET: Den amerikanska rollen i kriget 1973 har missuppfattats i vida kretsar.

**Example 27:**

SOURCE: the new data type

TARGET: den nya datatypen

In the following example the definite article (or the determinative pronoun) is present but the head word is not definite which means that the corresponding unit for *the* will only be the marker of definiteness in Swedish, namely *det*.

**Example 28:**

SOURCE: in the yellow house which we bought last year.

TARGET: i det gula hus som vi köpte förra året

***NP/ARTICLE RULE 2:** If a noun phrase contains modifiers or classifiers and these correspond independently, then only the head word and possible article will be selected for correspondence.*

The NP/Article Rule 3 can be seen as a complement to the Proper Names Rule 2. Here are some examples:

**Example 29:**

SOURCE: displays **the** Quantity **field**

TARGET: visar **fältet** Antal

**Example 30:**

SOURCE: which you enter in **the** Required Date **field**

TARGET: som skrivs in i **fältet** Begärt leveransdatum

**Example 31:**

SOURCE: Microsoft Access displays **the** Export **dialog box**.

TARGET: **Dialogrutan** Exportera visas.

**Example 32:**

SOURCE: We were advised to go on foot, along the old **State Department Building**

TARGET: Man rådde oss att gå till fots, längs **utrikesdepartementets gamla byggnad**

**Example 33:**

SOURCE: click the **Datasheet View** button on the toolbar

TARGET: klicka på **Datablad** i verktygsfältet

In the above example the classifier “button” has been omitted in the translation.

***NP/ARTICLE RULE 3:** Possessive pronouns in English can correspond to definite noun phrases in Swedish. For example: “your computer” – “datorn”, “his head” – “huvudet”.*

**Example 34:**

SOURCE: after you run **your** query

TARGET: när du har kört **frågan**

## 4.5 Verb constructions and infinitive markers

Using the two rules of thumb presented in the beginning of section 3, the handling of verb constructions with auxiliaries is relatively straightforward.

To specify the handling of verb constructions, we can formulate the following rule:

**VERB CONSTRUCTION RULE 1:** *If an auxiliary on the source side corresponds to an auxiliary on the target side, then the pair should be annotated as a link.*

**Example 35:**

SOURCE: The bathing cabins were nailed shut.

TARGET: Badhytterna var igenspikade.

**Example 36:**

SOURCE: Of course, theoretically you should n't have any orders without customers.

TARGET: I verkligheten ska det naturligtvis inte finnas några beställningar utan kunder.

**Example 37:**

SOURCE: if you don't want to be prompted each time ...

TARGET: om du inte vill bli ombedd varje gång ...

In other cases, the procedure is the standard one, namely to follow the rules of thumb, to find the longest and at the same time the shortest correspondences in the text.

**Example 38:**

SOURCE: Dates and times are displayed this way.

TARGET: Datum och tidsformat visas så här.

**Example 39:**

SOURCE: the data types you want

TARGET: de data som du vill ha

**Example 40:**

SOURCE: if you 're attaching an SQL table ...

TARGET: när du kopplar en SQL-tabell ...

**Example 41:**

SOURCE: Microsoft Access displays a message telling you how many records are to be deleted.

TARGET: Ett meddelande visas om hur många poster som kommer att tas bort.

In the normal case, the English infinitive marker “to” is treated under the rules of thumb, that is, that it is regarded as a single unit, and should consequently be selected to correspond with the infinitival “att” in Swedish. If there is no infinitive marker in the target, the source “to” should be categorized as a “null link”.

There is however an exception to the application of the rules of thumb described above. This concerns the group of modal verbs in English that are closely connected to the infinitival “to”, such as “need”, “want” and “ought”. The handling of these can be stated as follows:

**VERB CONSTRUCTION RULE 2:** *If the English verb is closely attached to the infinitive marker “to”, then the verb and the infinitive marker “to” should be regarded as one unit.*

**Example 42:**

SOURCE: if others **need to** use it

TARGET: om andra **behöver** använda tabellen

A final case for verb construction regards the use of clauses with finite verbs and their correspondence to clauses with infinite verbs. In such cases there is often a difference regarding the presence of explicit subjects, which is illustrated in the following example:

**Example 43:**

SOURCE: ... just as you can when **using** a database exclusively

TARGET: ... på samma sätt som när du **använder** en databas exklusivt

Here it is sufficient to mark the infinitival “using” with the finite form “använder”, thus the target subject “you” is regarded as a kind of grammatically necessary addition in the target sentence.

The “do”-support in English needs special treatment. The stand we take here is to ignore the different forms of “do” when it is used as an auxiliary, which can be illustrated with the following examples:

**Example 44:**

SOURCE: I **do** not know how much longer I’ll need to stay ...

TARGET: Jag vet inte hur pass mycket längre jag till jag behöver stanna ... (*Null link*)

**Example 45:**

SOURCE: ... where **did** that come from?

TARGET: ... varifrån hade den kommit? (*Null link*)

**Example 46:**

SOURCE: ... he does this every day

TARGET: ... han **gör** så här jämt (*Regular link*)

Particle verbs should always be marked as whole units, that is, both the verb and the particle should be selected.

**Example 47:**

SOURCE: ... the plane **took off** at 3 p.m.

TARGET: ... planet **lyfte** kl 15. (*Regular link*)

If there are particle verbs on both the source and the target side, the whole verb constructions should be selected. We regard particle verbs as lexical items (just like proper names and lexicalized collocations).

**Example 48:**

SOURCE: ... he **gave up** at the end of the sixth round.

TARGET: ... han **gav upp** i slutet av sjätte ronden (*Regular link*)

## 4.6 Paraphrases

As described in section 4.1, sometimes it is difficult to distinguish between paraphrases and omissions. However, Omission Rule 1 should give guidelines when determining whether a certain link should be seen as a paraphrase or an omission.

Paraphrases will often be marked as “fuzzy links”, if the relationship between the source and target unit is approximate. Some examples of links that can be found in paraphrases and which also are judged as “fuzzy”.

**Example 49:**

SOURCE: to add **more than one** table at a time

TARGET: om du vill lägga till **flera** tabeller på en gång (*fuzzy link*)

**Example 50:**

SOURCE: For more information, search Help for “Count function”.

TARGET: Mer information **finns i** Hjälp under “Antal, funktion”.  
(*fuzzy link*)

**Example 51:**

SOURCE: if you do n't want any fields on the form to be modified ...

TARGET: om du vill att **inga** fält ska kunna ändras ... (*fuzzy link*)

## 4.7 Definite expressions and pronouns

It is relatively obvious that there could be correspondences between definite descriptions and pronouns in a translation. We can define the following rule to cover this relationship:

**PRONOUN RULE 1:** *If there is a unit on one side which expresses some kind of anaphoric or deictic reference type and a definite description on the other side, then mark the two units for correspondence but as a fuzzy link.*

Two examples will illustrate the above rule:

**Example 52:**

SOURCE: repeat **this procedure** until all the fields you want ...

TARGET: upprepa **enligt ovan** till dess att alla fält som du vill ta med ... (*fuzzy link*)

**Example 53:**

SOURCE: The property sheet defines the characteristics of the control, such as its name, the source of its data, and its format.

TARGET: I egenskapsfönstret definieras kontrollens egenskaper, t ex **kontrollens** namn, datakälla och format. (*fuzzy link*)

## 5. ACKNOWLEDGEMENTS

Many thanks to Lars Ahrenberg, Anna Sångvall Hein, Jörg Tiedemann and Mikael Andersson for valuable comments on earlier drafts of this document.

## 6. 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 17<sup>th</sup> International Conference on Computational Linguistics*, Montréal, Canada, 10-14 August 1998, 29-35.

- Merkel M. & L. Ahrenberg (1998) *Evaluating Word Alignment Systems*. Unpublished PLUG report, Linköping university.
- Melamed, I. D. (1998). *Annotation Style Guide for the Blinker Project*. , IRCS Technical Report #98-06, University of Pennsylvania.
- Véronis, J. & P. Langlais (1999). "Evaluation of parallel text alignment systems". In *Parallel Text Processing* (ed. J. Véronis), Kluwer (this volume).