Annotation Guidelines for Dutch-English Word Alignment

version 1.0

LT3 Technical Report – LT3 10-01

Lieve Macken

LT3 – Language and Translation Technology Team
Faculty of Translation Studies
University College Ghent
URL: http://veto.hogent.be/lt3[1]

April 22, 2010

[1]The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from http://veto.hogent.be/lt3/publications_en.html.

Contents

Chapter 1

Introduction

The goal of the annotation task is the creation of a reference alignment for a set of English-Dutch parallel texts. Manually created reference alignments – also called Gold Standards – have been used to develop or test automatic word alignment systems (Melamed, 2001b; Véronis, 2000).

As translations are characterized by both correspondences and changes, three types of links are introduced: regular links are used to connect straightforward correspondences, fuzzy links for translation-specific shifts of various kinds (paraphrases and divergent translations), and null links for source text units that have not been translated or target text units that have been added.

This annotation style guide is to a large extent based on the annotation guidelines of other word alignment projects (Melamed, 2000; Merkel, 1999; Och and Ney, 2003; Véronis, 1998). As a starting point, the Blinker project (Melamed, 2001a) was used, because of the identical nature of the annotation task. The Blinker project aimed at aligning all words between two parallel texts. The aim of the Arcade project (Véronis, 1998) and the Plug project (Merkel, 1999) was translation spotting: only for some given words was the translation in the target text selected. However, useful elements of the Arcade and Plug guidelines were incorporated in these guidelines, e.g. the distinction between regular and divergent translations, which is reflected in regular and fuzzy links.

To make the manual annotations as useful as possible for different types of alignment projects, a multi-level annotation is proposed in case of divergent translations: fuzzy links are used to connect paraphrased sections, regular links are used to connect corresponding words within the paraphrased sections.

When comparing the four above-mentioned guidelines, most disagreement was found in the rules covering function words (determiners, auxiliaries and prepositions and the like). We have tried to come up with consistent rules to link function words that have no direct counterpart in the other language.

The guidelines have also been adapted for the language pair English-Dutch, and contain some rules to describe language-specific phenomena.

This document consists of two sections: general guidelines and detailed guidelines. The detailed section contains rules for the annotation of noun phrases, verbal constructions, adverbials, referring expressions, punctuation etc. The detailed guidelines can be seen as a language-specific implementation of the general guidelines.

HandAlign[1] is used as annotation tool. The screenshots in this document are taken from the alignment window of the HandAlign annotation tool.

As in most manual annotation projects, these guidelines are not final. This document will be updated regularly in the course of the annotation process.

[1] http://www.cs.utah.edu/~hal/HandAlign/

Chapter 2

General Guidelines

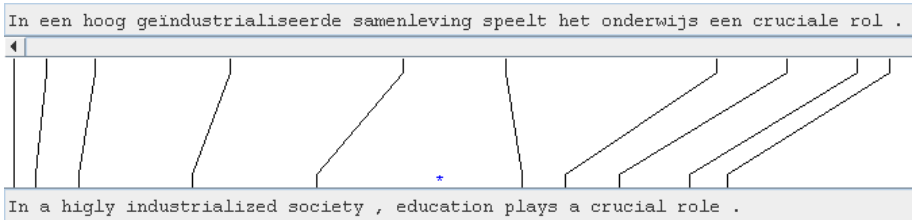
The annotators will be working with Dutch and English texts (sentences, paragraphs or complete texts) that are translations of each other. The corpus to be annotated is bidirectional and contains Dutch text translated into English as well as English texts translated into Dutch. The task of the annotator is to identify all correspondences in the source and target sentences.

The annotators will be asked to indicate the *minimal* language unit in the source text that corresponds to an equivalent in the target text[1], and vice versa.

To determine this minimal language unit, two major rules can be formulated (Merkel, 1999; Véronis, 1998):

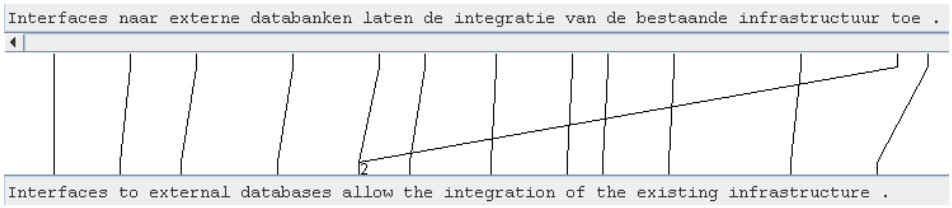
1. Select as many words as necessary in the source and in the target sentence to ensure a two-way equivalence
2. Select as few words as possible in the source and in the target sentence, while preserving two-way equivalence

In the first example there is word-by-word correspondence for all words except for *het onderwijs* ∼ *education*: in the English sentence there is no definite article.



[1]Cf. Barkhudarov's definition of translation unit (Barkhudarov, 1993, p. 40): a unit in the source text for which an equivalent can be found in the text of the translation but whose elements, taken separately, do not correspond to equivalents in the translated text.

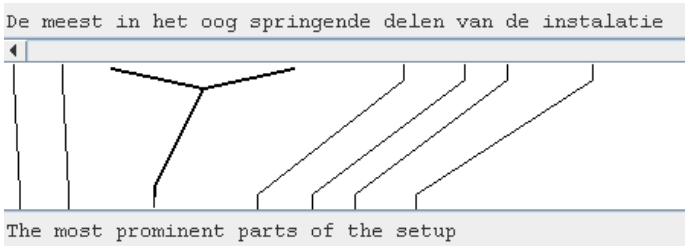
The corresponding units are not necessarily contiguous[2], e.g. *laten...toe* ∼ *allow*.



In most translations however, translational correspondences are more complex, and only for some words, word-by-word correspondences can be found. The rest of the sentence is translated on the level of combination of words.

2.1 Phraseological units

One example of translation on the level of combination of words is the translation of phraseological units. Phraseological units can be compounds, idioms, fixed expressions, multiword abbreviations, proper names, specific terms, and the like. In most cases, the meaning of a phraseological unit cannot be derived from the (literal) meaning of its parts. Phraseological units have to be treated as single units on both the source and target side, e.g. *in het oog springend* ∼ *prominent*, *deel uitmaken van* ∼ *to be part of*.



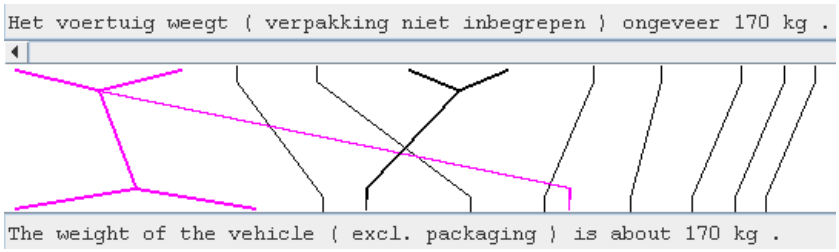
[2]The number '2' in the screenshot indicates that there are two words aligned to *allow*. In case of one-to-many or many-to-one links, the number of links is printed in the alignment window.

2.2 Paraphrases and divergent translations

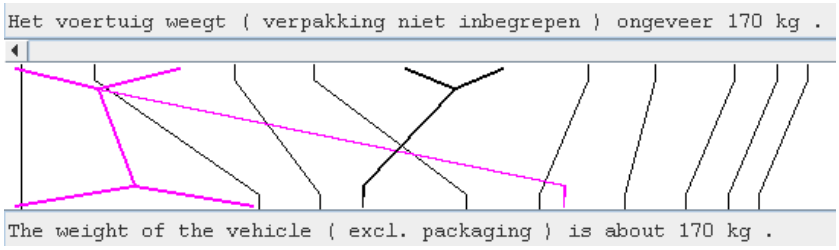
In some cases, translational correspondence cannot be indicated at the level of words or word groups (with the same constituent structure) as the translator has completely rephrased the fragment. In these cases the whole phrase should be selected and marked as a fuzzy link (1). In HandAlign, fuzzy links are drawn in magenta. Regular links are drawn in black.

If some words or word groups within the paraphrased section clearly correspond, mark these with a regular link (2).

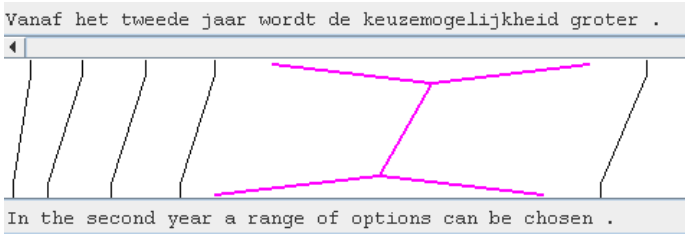
(1) Fuzzy link: *het voertuig weegt ~ the weight of the vehicle ... is*



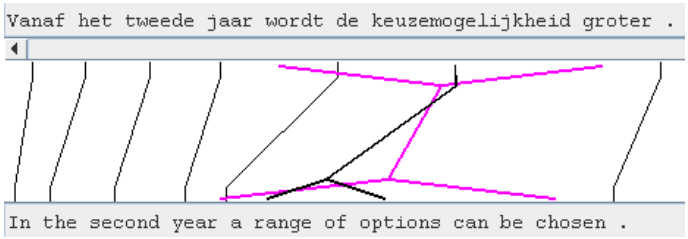
(2) Regular links: *Het ~ the*, *voertuig ~ vehicle*



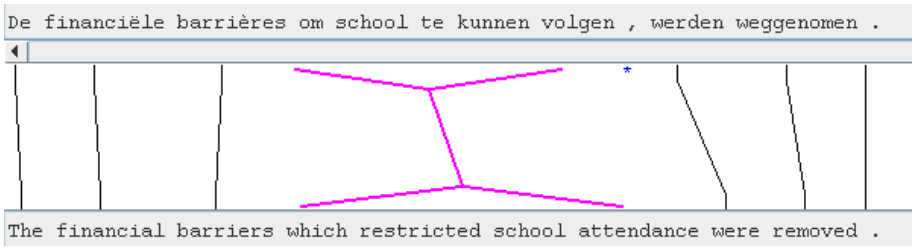
(1) Fuzzy link: *wordt de keuzemogelijkheid groter ~ a range of options can be chosen*



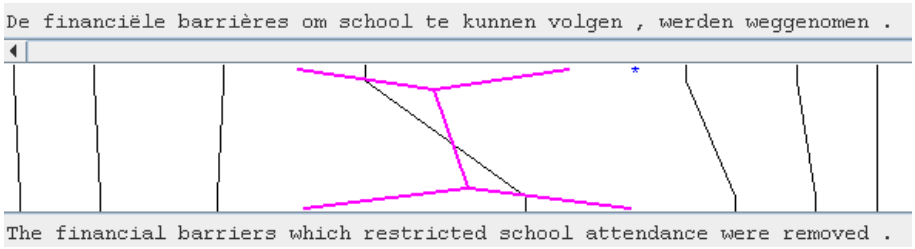
(2) Regular links: *de ~ a*, *keuzemogelijkheid ~ range of options*



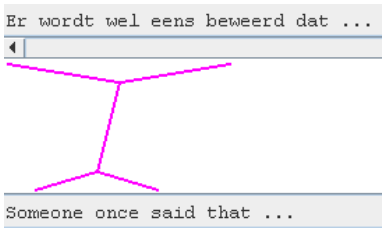
(1) Fuzzy link: *om school te kunnen volgen ∼ which restricted school attendance*



(2) Regular link: school ∼ school

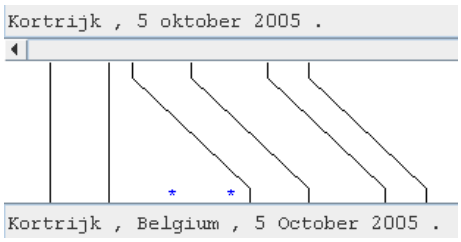


(1) Fuzzy link: *Er wordt wel eens beweerd ∼ Someone once said*



2.3 Omissions

In the translation process, the translator may have omitted or inserted some words. Words whose meaning is not expressed in the other language (either source or target language) should be indicated as null link. Null links are visualized by an asterisk.



2.4 Summary

Different language units can be linked: words, punctuation marks, word groups or paraphrased sections. Three types of links are used: regular, fuzzy or null links.

Regular link:	Similar meaning (semantically equivalent) and similar constituent structure or identical syntactic role.
Fuzzy link:	Semantically overlapping; similar meaning but different structure (other perspective, different part of speech, different syntactic role, ...). Fuzzy links are also used to connect different types of phrase, e.g. prepositional phrases to noun phrases, e.g. in adverbials, adnominals, indirect objects.
Null link:	Meaning not expressed / no formal equivalent in the other language. By definition, null links can only be used for content words or word groups containing at least one content word.

A multi-level annotation is used in case of fuzzy links. If some words or word groups within a fuzzy link clearly correspond, mark these with a regular link. It is not necessary to mark null links within fuzzy links.

The multi-level annotation scheme is only used for regular links within fuzzy links. Do not use regular links within regular links.

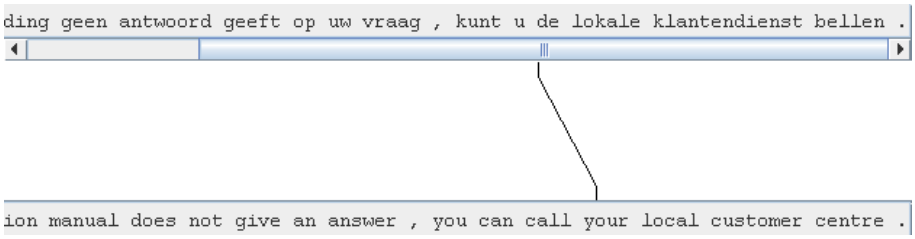
Chapter 3

Detailed Guidelines

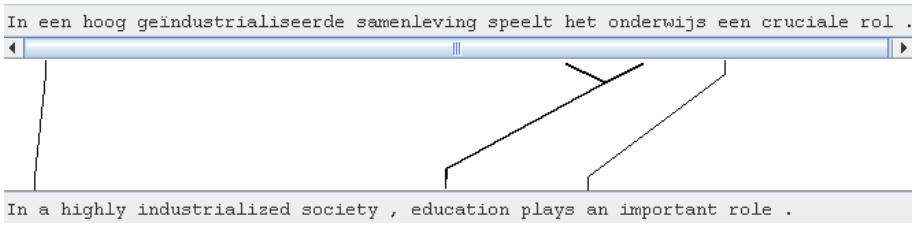
3.1 Noun Phrases

3.1.1 Determiners

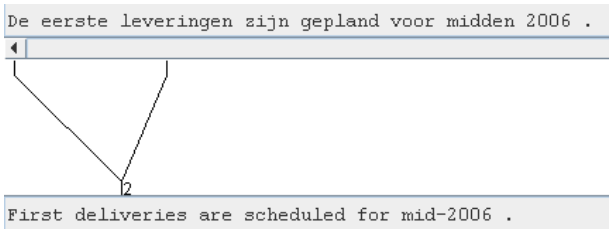
Determiners can be connected with a regular link, regardless whether they are articles or possessive pronouns.



Extra determiners in source or target language should be linked together with their noun to the noun's translation with a regular link.



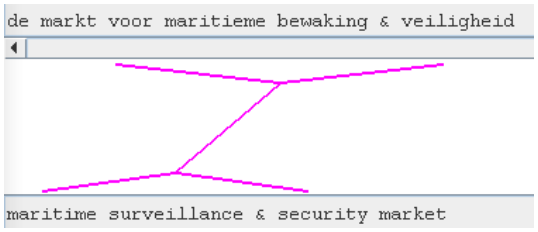
Do not include modifiers when linking a determiner together with the noun to the noun's translation (i.e. the determiner and the noun are not necessarily contiguous).



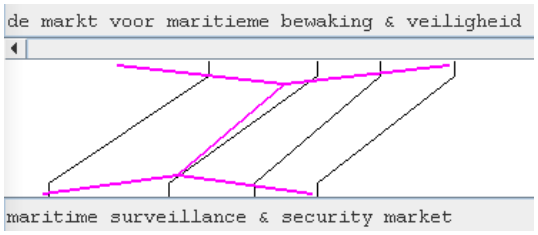
3.1.2 Pre- vs. post-modifiers

English pre-modifiers often correspond with Dutch post-modifiers. Use a fuzzy link to connect the complete pre-modifier with the post-modifier (1). Use regular links to connect corresponding words within the modifiers (2).

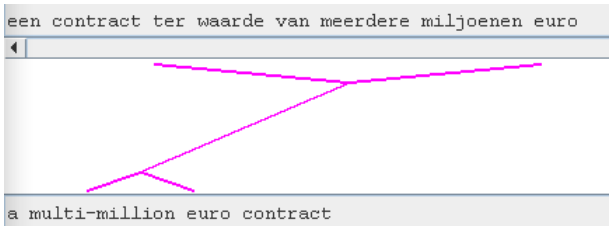
(1) Fuzzy link: *voor maritieme bewaking & veiligheid* ∼ *maritime surveillance & security*



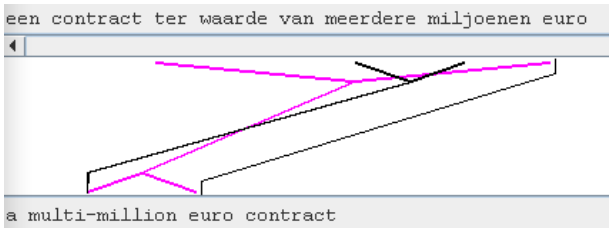
(2) Regular links: *maritieme* ∼ *maritime*, *bewaking* ∼ *surveillance*, *&* ∼ *&*, *veiligheid* ∼ *security*



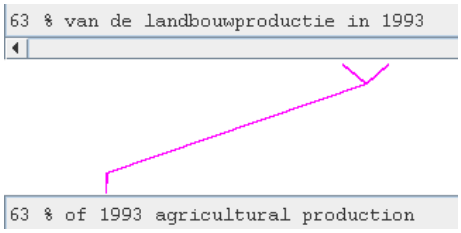
(1) Fuzzy link: *ter waarde van meerdere miljoenen euro* ∼ *multi-million euro*



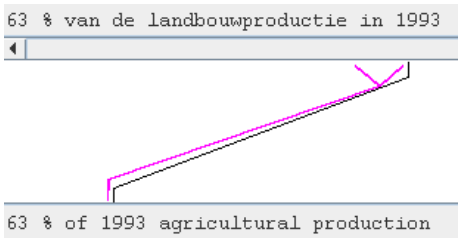
(2) Regular links: *meerdere miljoenen* ∼ *multi-million*, *euro* ∼ *euro*



(1) Fuzzy link: *in 1993* ∼ *1993*

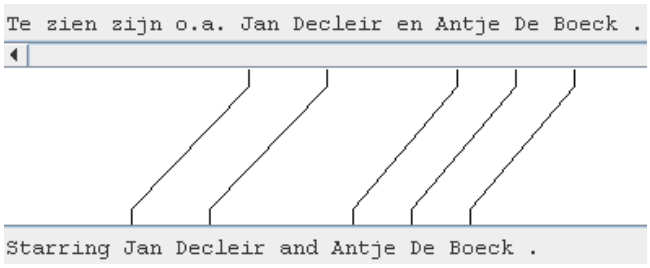


(2) Regular link: *1993* ∼ *1993*



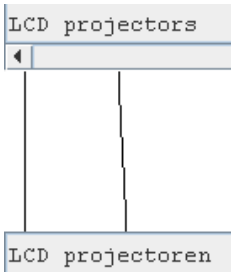
3.1.3 Proper names

Link the corresponding parts of multi-word proper names by means of a regular link.

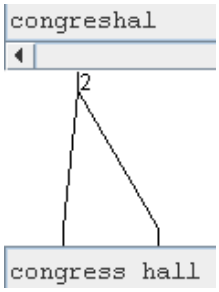


3.1.4 Compounds

Link the corresponding subparts of the compounds by means of a regular link.



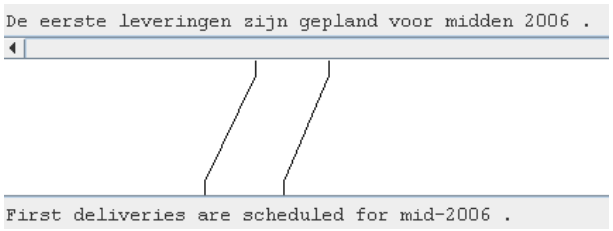
Use multiple regular links if the English compound is a multiword and the Dutch compound is single word.

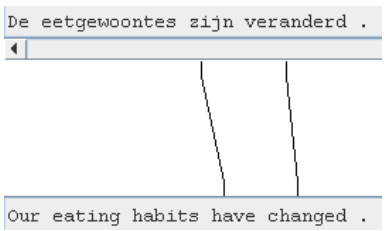


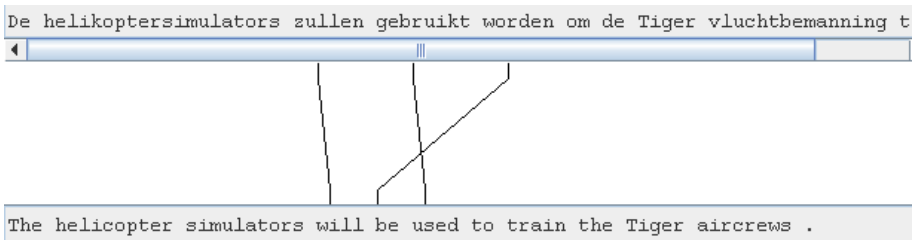
3.2 Verb Phrases

3.2.1 Auxiliary verbs

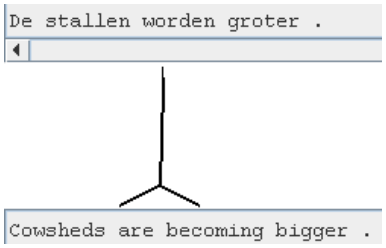
If an auxiliary in the source sentence has a corresponding auxiliary in the target sentence, use a regular link to connect the auxiliaries.

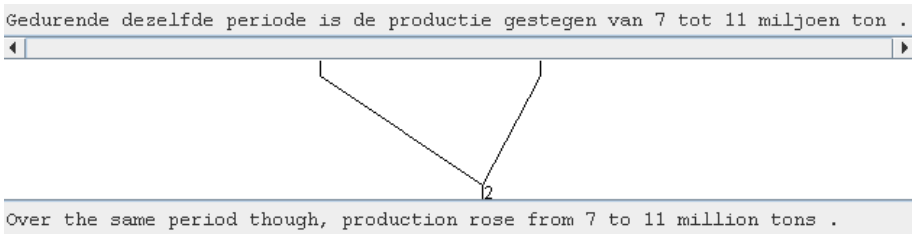




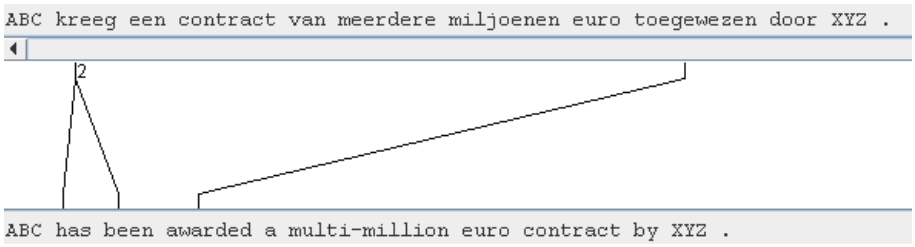


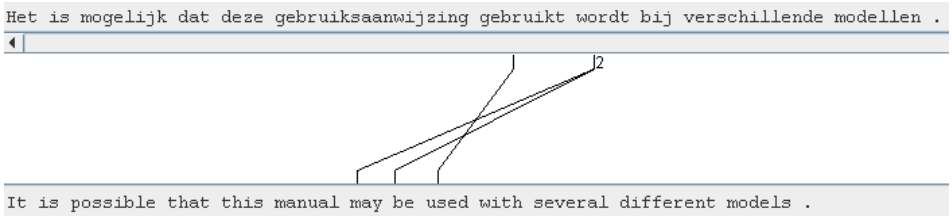
If the main verb of one language has no auxiliaries attached, connect the auxiliaries in the other language together with the main verb to the verb's translation with a regular link. In case of active-passive transformation use a fuzzy link (see "active vs. passive constructions").





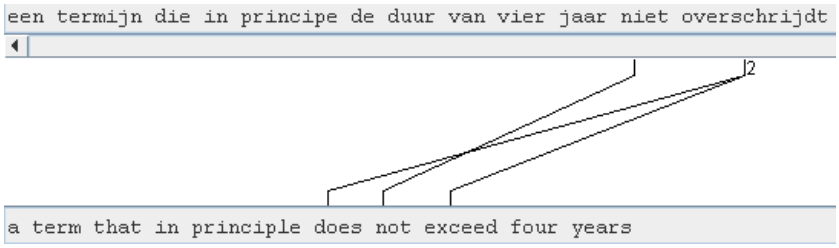
If more auxiliaries are attached to the main verb in one language, group the auxiliaries and connect them with the corresponding auxiliary with a regular link.

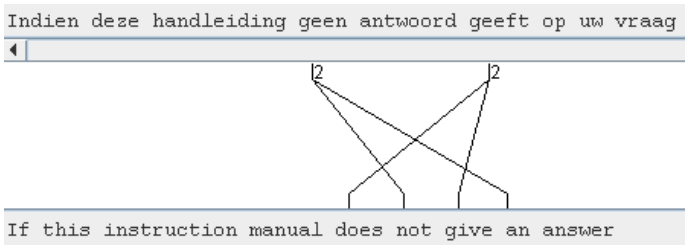




3.2.2 Negation and do-support

Link the auxiliary "do" together with the main verb to the verb's translation with a regular link.

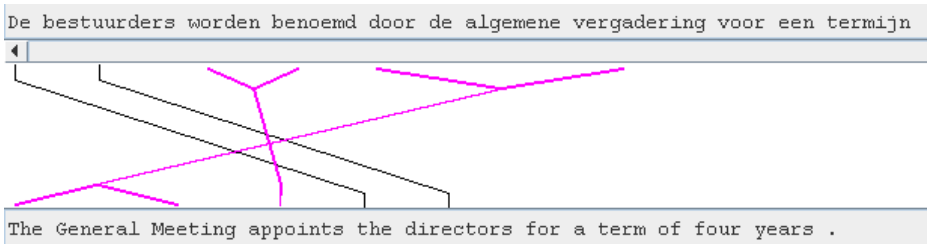




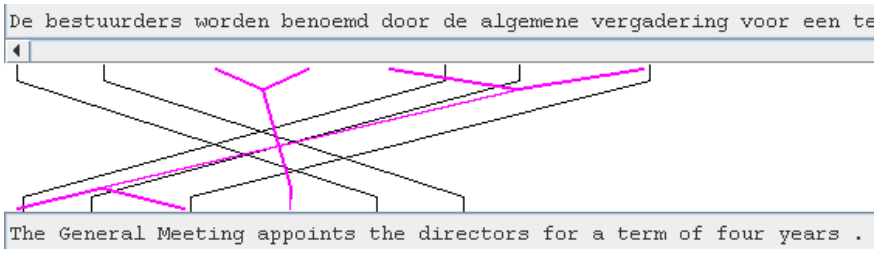
3.2.3 Active vs. passive constructions

If an active construction is translated by a passive construction or vice versa, use fuzzy links to connect the corresponding verbs and the corresponding agents (1). Use regular links to connect corresponding words within the agent (2).

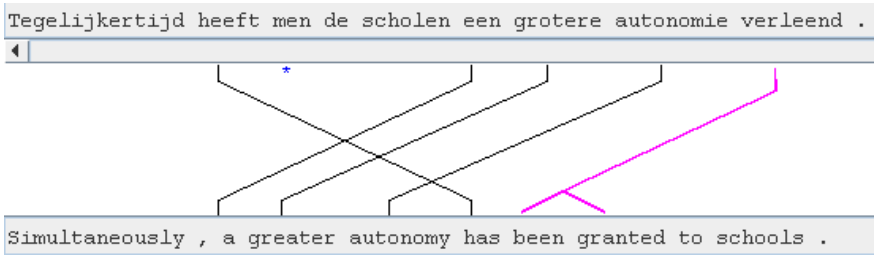
(1) Fuzzy links: *worden benoemd* ~ *appoints*, *door de algemene vergadering* ~ *the General Meeting*



(2) Regular links: *de* ~ *the*, *algemene* ~ *General*, *vergadering* ~ *Meeting*

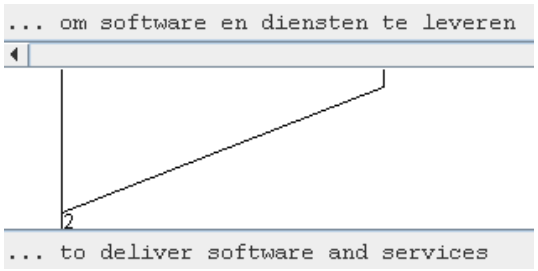


Use a null link to mark the agent of the active sentence that is not expressed in the passive translation.

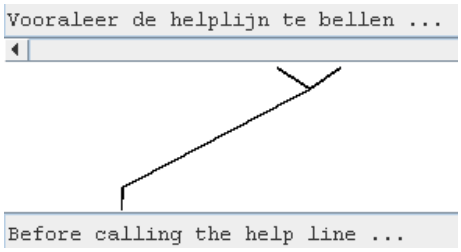


3.2.4 Infinitive marker “te”

If the Dutch construction “om ... te” corresponds with English “to” use a regular link to connect “om ... te” to “to”.



In other cases, connect the Dutch infinitive marker “te” (without “om”) together with the infinitive to the infinitive’s translation with a regular link.

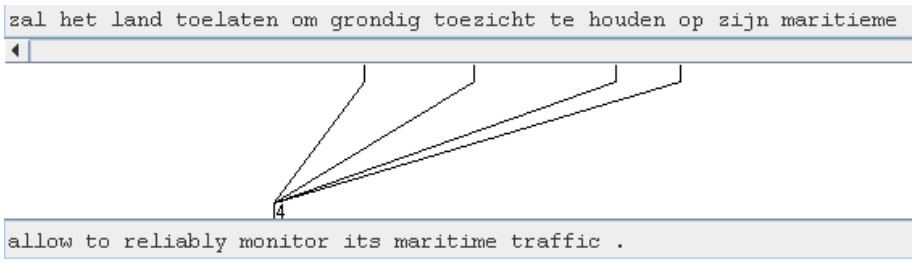


3.2.5 Phrasal verbs

Consider phrasal verbs as one lexical unit. Connect particles that are part of a phrasal verb together with the verb to the phrasal verb’s translation with a regular link.

3.2.6 Verb complementation

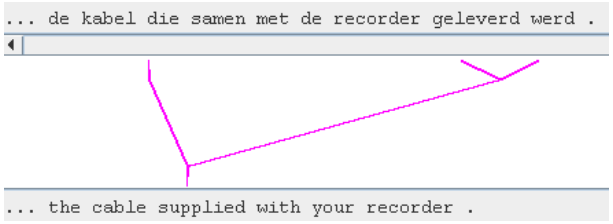
If a verb requires a prepositional phrase as its complement, and the verb's translation a noun phrase or vice versa, connect the preposition of the prepositional phrase together with the verb or verbal group to the verb's translation with a regular link.



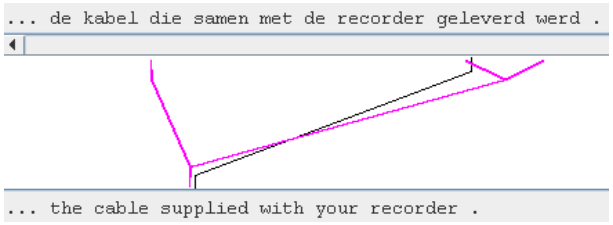
3.2.7 Participles vs. relative clauses

If a participle is translated by a relative clause, connect the relative pronoun together with the verb of the relative clause to the participle with a fuzzy link (1). Connect the corresponding verbs with a regular link (2).

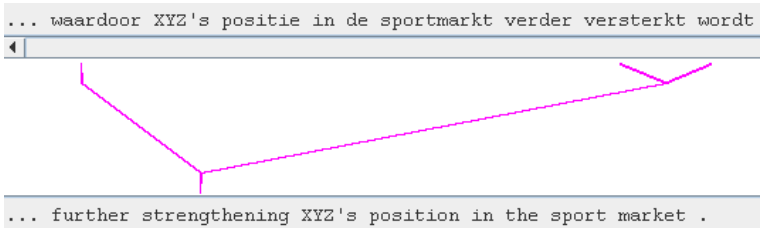
(1) Fuzzy link: *die ... geleverd werd* ∼ *supplied*



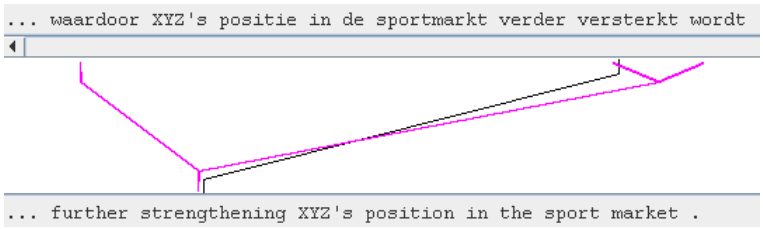
(2) Regular link: *geleverd* ∼ *supplied*



(1) Fuzzy link: *waardoor ... versterkt wordt* ∼ *strengthening*



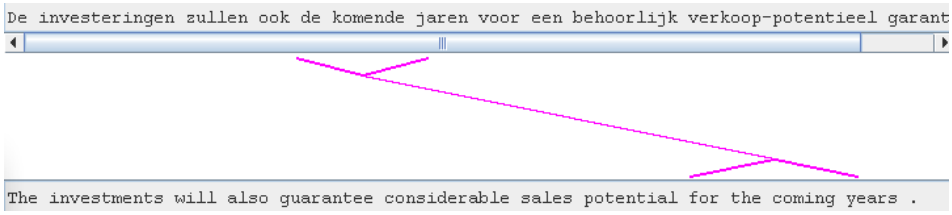
(2) Regular link: *versterkt* ∼ *strengthening*



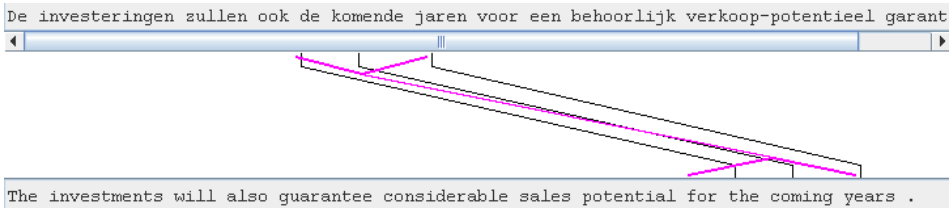
3.3 Noun Phrases vs. Prepositional Phrases

If a prepositional phrase corresponds to a noun phrase or vice versa (e.g. in adverbials, adnominals, indirect objects), use a fuzzy link to connect the corresponding phrases. Use regular links to connect the corresponding words within the phrases.

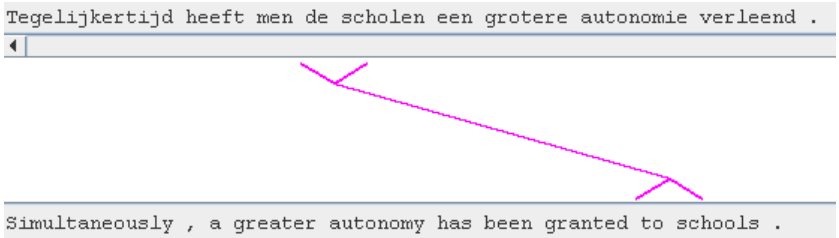
(1) Fuzzy link: *de komende jaren* ∼ *for the coming years*



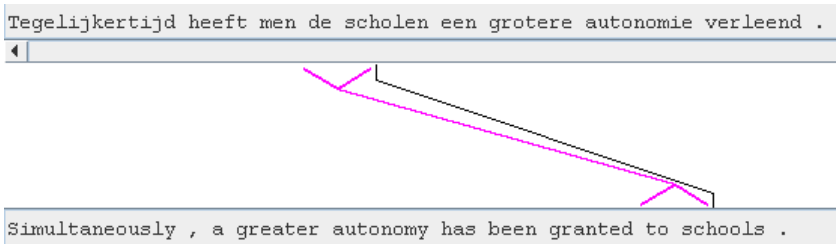
(2) Regular links: *de* ∼ *the*, *komende* ∼ *coming*, *jaren* ∼ *years*



(1) Fuzzy link: *de scholen* ∼ *to schools*

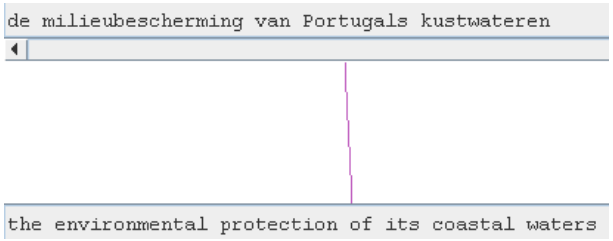


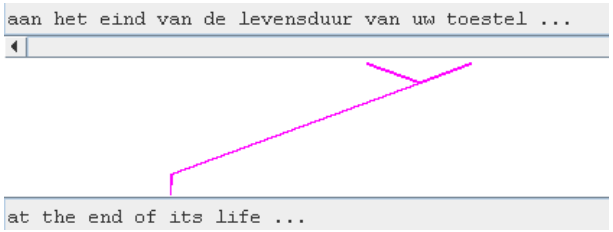
(2) Regular link: *scholen* ∼ *schools*



3.4 Referring expressions

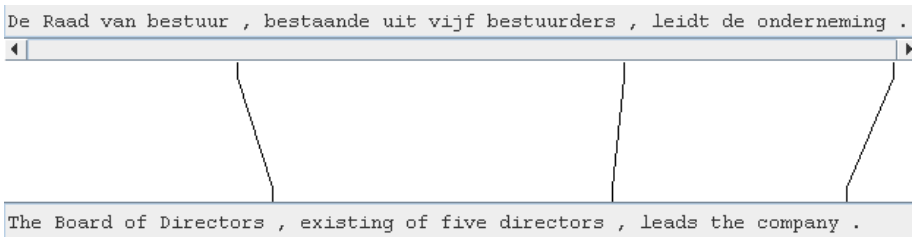
If a pronoun or another referring expression corresponds with a definite description, use a fuzzy link.



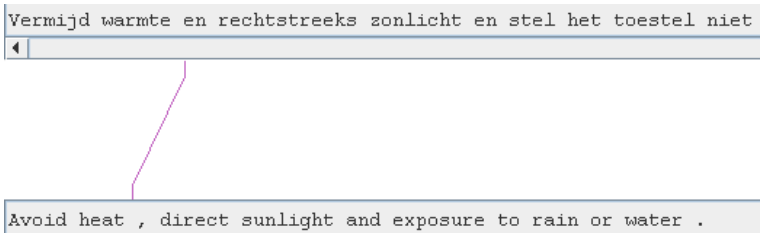


3.5 Punctuation

Connect corresponding punctuation marks with a regular link.

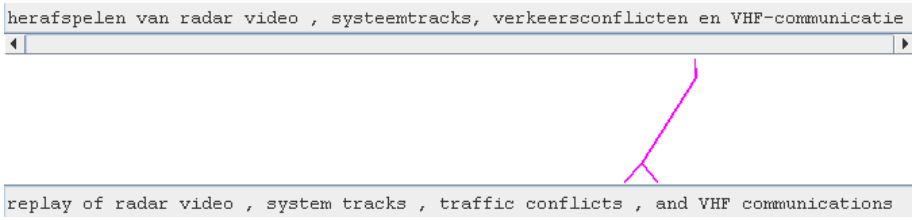


If a punctuation mark corresponds to a word or to another type of punctuation mark, use a fuzzy link.

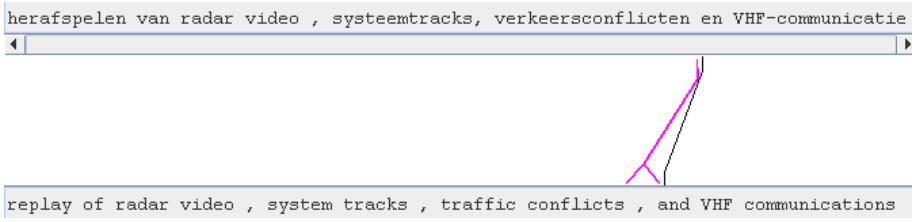


If a conjunction is expressed by a comma and a conjunction in one language and only by a conjunction in the other language, connect both the comma and the conjunction to the conjunction with a fuzzy link (1). Use a regular link to connect the corresponding conjunctions (2).

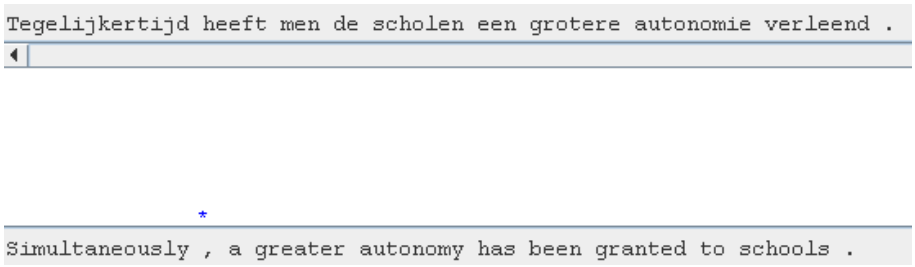
(1) Fuzzy link: *en* ∼ *, and*



(2) Regular link: *en* ∼ *and*

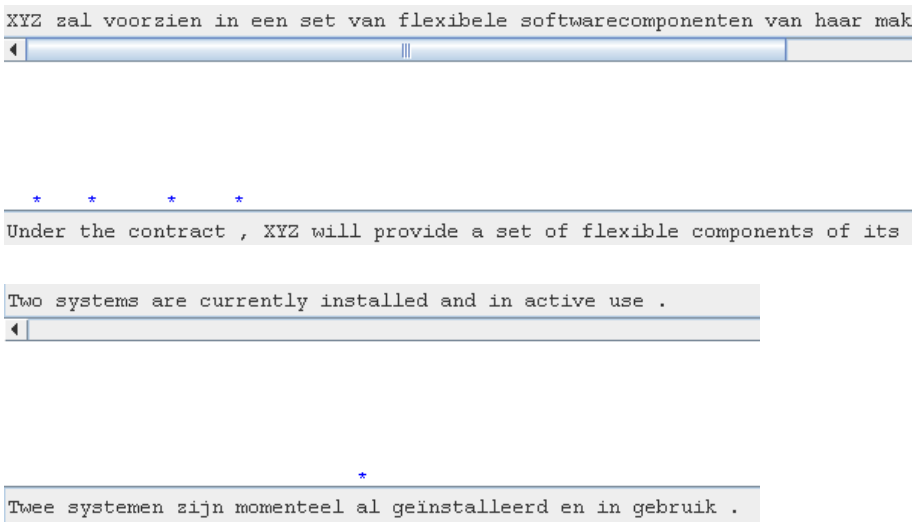


If a punctuation mark cannot be linked, mark it as an omission.



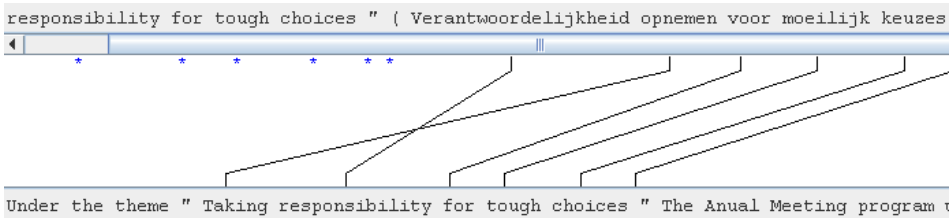
3.6 Omissions

Use a null link to mark words whose meaning is not expressed in the source or target language. Null links are visualized by an asterisk.



3.6.1 Non-translated segments

If the translator has inserted both a non-translated phrase and its translation, mark the non-translated phrase with a null link.

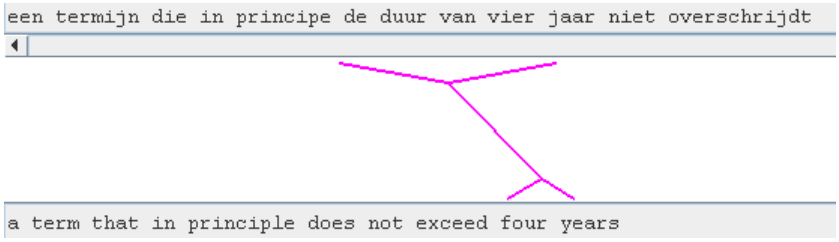


3.6.2 Omissions vs. paraphrases

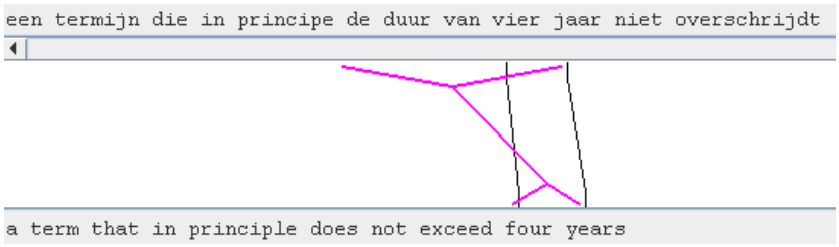
Use null links only for words whose meaning is not expressed in the other language. If a meaning is paraphrased or expressed more explicitly in source or target sentence, use a fuzzy link (1).

If some words or word groups within the paraphrased section clearly correspond, mark these with a regular link (2).

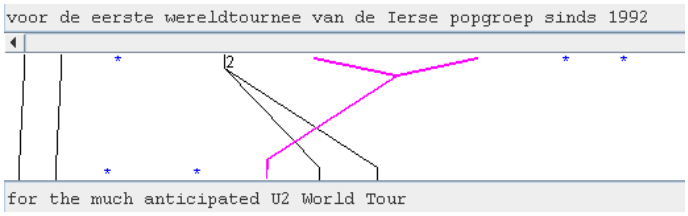
(1) Fuzzy link: *de duur van vier jaar* ~ *four years*



(2) Regular link: *vier jaar* ~ *four years*



Use null links in source and target language, if a phrase is paraphrased in such a way that some words in both source and target language are not expressed in the other language.



3.7 Quick reference guide

	Correspondence scope	Type of link
Determiners	determiner ↔ determiner	regular
	determiner + noun ↔ noun	regular
Premodification vs. postmodification	(1) premodifier ↔ postmodifier	fuzzy
	(2) corresponding words	regular
Auxiliaries	auxiliaries ↔ auxiliaries	regular
	auxiliaries + verb form ↔ verb form	regular
Active vs. passive constructions	auxiliaries + passive verb form ↔ active verb form	fuzzy
	(1) agent of active construction ↔ agent of passive construction	fuzzy
	(2) corresponding words of agent	regular
Infinitive marker "te"	"om ... te" ↔ "to"	regular
	"te" + verb form ↔ verb form	regular
Phrasal verbs	verb + particle ↔ verb form	regular
Participles vs. relative clauses	(1) participle ↔ relative pronoun + verb	fuzzy
	(2) corresponding verbs	regular
Noun phrases vs. prepositional phrases	(1) noun phrase ↔ prepositional phrase	fuzzy
	(2) corresponding words	regular
Referring expressions	referring expression ↔ definite description	fuzzy
Punctuation	punctuation mark ↔ identical punctuation mark	regular
	punctuation mark ↔ different punctuation mark	fuzzy
	(1) punctuation mark + conjunction ↔ conjunction	fuzzy
	(2) conjunction ↔ conjunction	regular
Paraphrases	(1) paraphrased section ↔ paraphrased section	fuzzy
	(2) corresponding words	regular

References

Barkhudarov, Leonid. 1993. The problem of the unit of translation. In P. Zlateva, editor, *Translation as social action: Russian and Bulgarian perspectives*. Routledge, London, pages 39–46.

Melamed, Dan I. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Melamed, Dan I. 2001a. Annotation style guide for the Blinker Project. In Dan I. Melamed, editor, *Empirical methods for exploiting parallel texts*. MIT Press, Cambridge, Massachusetts, pages 169–182.

Melamed, Dan I. 2001b. Manual annotation of translational equivalence. In Dan I. Melamed, editor, *Empirical methods for exploiting parallel texts*. MIT Press, Cambridge, Massachusetts, pages 65–77.

Merkel, Magnus. 1999. Annotation Style Guide for the PLUG Link Annotator.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Véronis, Jean. 1998. Arcade. Tagging guidelines for word alignment. Version 1.0.

Véronis, Jean. 2000. Evaluation of parallel text alignment systems: the ARCADE project. In Jean Véronis, editor, *Parallel text processing: alignment and use of translation corpora*. Kluwer Academic Publishers, Dordrecht, pages 369–388.