An annotation scheme and Gold Standard for Dutch-English word alignment

Lieve Macken

Language and Translation Technology Team, University College Ghent, Belgium
Dept of Applied Mathematics and Computer Science, Ghent University, Belgium
Lieve.Macken@hogent.be

Abstract

The importance of sentence-aligned parallel corpora has been widely acknowledged. Reference corpora in which sub-sentential translational correspondences are indicated manually are more labour-intensive to create, and hence less wide-spread. Such manually created reference alignments – also called Gold Standards – have been used in research projects to develop or test automatic word alignment systems. In most translations, translational correspondences are rather complex; for example word-by-word correspondences can be found only for a limited number of words. A reference corpus in which those complex translational correspondences are aligned manually is therefore also a useful resource for the development of translation tools and for translation studies. In this paper, we describe how we created a Gold Standard for the Dutch-English language pair. We present the annotation scheme, annotation guidelines, annotation tool and inter-annotator results. To cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles, our Gold Standard data set contains texts from different text types. The Gold Standard will be publicly available as part of the Dutch Parallel Corpus.

1. Introduction

Reference corpora in which sub-sentential translational correspondences are indicated manually – also called Gold Standards – have been used as an objective means for testing word alignment systems (Melamed, 1998; Och and Ney, 2003). In most translations, translational correspondences are rather complex; for example word-by-word correspondences can be found only for a limited number of words. A reference corpus in which those complex translational correspondences are aligned manually is therefore also a useful resource for benchmarking sub-sentential translation memory systems (Macken, 2009) and for studying shifts of translation, the linguistic changes that occur in the process of translating (Bakker et al., 2008).

In this paper, we describe how we created a Gold Standard for the Dutch-English language pair. To cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles, our Gold Standard data set contains texts from different text types.

The Gold Standard will be publicly available as part of the Dutch Parallel Corpus (Macken et al., 2007; De Clercq and Montero Perez, 2010), which will be distributed by the Dutch Agency for Human Language Technologies (TST-centrale)[1]. In the Dutch Parallel Corpus project, a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French has been compiled. The DPC covers a broad range of text types and is balanced with respect to text type and translation direction.

2. Related annotation projects

Gold Standards of word alignments have been created for various language pairs, mainly to provide an objective way of evaluating word alignment systems (Melamed, 2001a; Och and Ney, 2003). However, there is no generally accepted standard method for creating such reference alignments.

A first important distinction that should be made is between *sample word* alignment and *full text* alignment. In the first case (e.g. the Arcade project (Véronis, 2000) and the PLUG project (Ahrenberg et al., 2002)), test words were selected on the basis of a number of criteria (e.g. frequency or polysemy characteristics) and only for these words translational correspondences are manually provided. In the second case, all words in the text are manually aligned.

It goes without saying that *full text* alignment is the more difficult task. Translational correspondences are often difficult to determine at the word level as in a sentence pair word-by-word correspondences can be found only for a limited number of words. The rest of the sentence is translated on the level of combinations of words. Two different approaches can be found in the literature to deal with those complex translational divergences in the manual annotation task.

In the first approach *ambiguous* alignments are explicitly allowed in the annotation scheme. Och and Ney (2003) introduced *sure* and *possible* links to create a reference set for English-French. Sure links were used for unambiguous alignments and possible links were used for ambiguous alignments (i.e. idiomatic expressions, free translations and missing function words). The approach was adopted by Lambert et al. (2005) for the English-Spanish language pair.

In the second approach, detailed annotation guidelines are used to provide clarity on how to align translational divergences. In the Blinker project, Melamed (2001a) created an elaborate annotation style guide for the French-English language pair. The reasonably high inter-annotator agreement rates show that the alignment task is feasible. The approach was adopted by Mihalcea and Pedersen (2003) for the Romanian-English test data of the HLT-NAACL 2003 workshop on building and using parallel texts.

Another respect in which annotation schemes differ is how they deal with null-alignments, i.e. source words that were not translated or target words that have been added during

[1] http://www.tst.inl.nl

translation. In some annotation projects (Melamed, 2001a; Mihalcea and Pedersen, 2003), the annotators were asked to explicitly mark those null-alignments, while in other projects (Och and Ney, 2003) all unlinked words were considered to be null-alignments.
In order to create a Gold Standard for English-Dutch, we opted for the second approach and defined detailed annotation rules. However, we do make a distinction between regular and divergent translations, which is reflected in the multi-level annotation scheme that is presented below.

3. Annotation scheme

In order to create an a-priori reference alignment for a set of English-Dutch parallel texts, translational correspondences were indicated manually. To that end an annotation scheme was created and detailed annotation guidelines were written.
To account for all the phenomena described above, three types of links were 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.
To make the manual annotations as useful as possible for different types of projects, a multi-level annotation is proposed in the case of divergent translations: fuzzy links are used to connect paraphrased sections, regular links are used to connect corresponding words within the paraphrased sections.
The main characteristics of the annotation scheme can be summarized as follows:

- All words are linked.
- Different units can be linked: words, word groups, punctuation marks, paraphrased sections.
- Discontinuous expressions can be linked.
- Three types of links are used: regular, fuzzy and null links.
- A multi-level annotation is used: regular links within fuzzy links are indicated.

4. Annotation guidelines

To improve consistency, detailed annotation rules (Macken, 2010) were written[2]. The annotation guidelines are to a large extent based on the annotation guidelines of other word alignment projects (Melamed, 2001a; Véronis, 1998; Merkel, 1999).
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. As explained above, the Arcade project (Véronis, 1998) and the Plug project (Merkel, 1999) were restricted to translation spotting: only for some given words the translational correspondence in the target text was indicated. However, useful elements of the Arcade and Plug guidelines were incorporated in our guidelines, e.g. the distinction between regular and divergent translations, which is reflected in regular and fuzzy links.
As a general rule, the minimal language unit in the source text that corresponds to an equivalent in the target text, and vice versa had to be aligned. To determine this minimal language unit, two major rules were taken from Véronis (1998) and Merkel (1999):

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

When comparing the three above-mentioned guidelines, most disagreement was found in the rules covering function words (determiners, auxiliaries, 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 Dutch-English language pair, and contain some rules to describe language pair-specific phenomena. The style guide 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, and the like, and can be seen as a language-specific implementation of the general guidelines.
Some example rules are given below:

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

5. Annotation Tool

To facilitate the annotation process, a graphical annotation tool, HandAlign[3], was used. The HandAlign annotation tool was originally developed for aligning articles and their summaries, but the tool offers enough flexibility to use it for other alignment purposes.
The annotator works in a graphical environment that consists of three panels (see figure 1):

- The top text area contains the source text.

[2] Available at http://veto.hogent.be/lt3/

[3] Available at http://www.cs.utah.edu/∼hal/HandAlign/

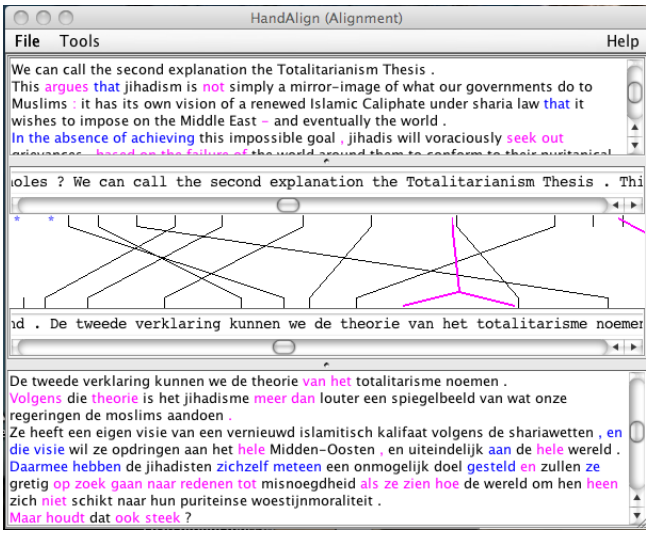

Figure 1: Graphical Annotation Tool

- The bottom text area contains the target text.
- The alignment area (in the middle) is where the source and target units can be selected and linked graphically.

In a preprocessing step, the input files were tokenized (punctuation was stripped off), and split into sentences. The HandAlign-tool offers the flexibility to define three types of links (regular links and fuzzy links are represented by another color; null links are represented by means of an asterisk) and to link multiword expressions and non-contiguous expressions.

For further processing, the output of HandAlign was converted into a table. In the table representation (see table 1), the first column contains the source text segment, the second column the target text segment, and the third column the type of link (F = fuzzy link, R = regular link). This table representation can be easily enriched with additional linguistic information (e.g. lemma and part-of-speech).

Source text segment	Target text segment	Type of link
We	we	R
can	kunnen	R
call	noemen	R
the	De	R
second	tweede	R
explanation	verklaring	R
the	de	R
Totalitarianism	van het totalitarisme	F
Totalitarianism	totalitarisme	R
Thesis	theorie	R
.	.	R

Table 1: Table representation of the manual alignments

6. Inter-annotator reliability

In order to assess the feasibility of the alignment task given the annotation guidelines, an inter-annotator experiment was set up. In a preliminary experiment, three staff members of the English department of the Faculty of Translation Studies of University College Ghent manually annotated eight texts of the corpus of press releases, amounting to a total of more than 10,000 words. The alignments of the staff members were compared with the author's alignments for those eight texts.

In order to familiarize the annotators with the annotation guidelines and the annotation tool, a training session with three training texts was organized. The training samples contained most of the examples of the annotation guidelines.

The annotators were asked to link all the words of the source and the target text. Null links were to be used for source text units that had not been translated or target text units that had been added. If the annotator forgot to link some words of source or target text, (s)he got a warning.

Translational equivalence is hard to establish for complex or divergent translations. Especially for such complex and divergent translations, comparing manual alignments is not a trivial task, as these alignments cannot be simply classified as *right* or *wrong*. Two different metrics were used to assess inter-annotator reliability: Kappa and Word Alignment Agreement.

6.1. Kappa

A widespread measure for evaluating inter-annotator agreement for tagging tasks in the field of computational linguistics is the Kappa statistic (Carletta, 1996; Di Eugenio and Glass, 2004). The Cohen's Kappa Statistic measures pairwise agreement among coders making category judgments. For a similar task, Daumé III and Marcu (2005) used the Kappa statistic to compute inter-annotator agreement for word-to-word and phrase-to-phrase alignments between abstract-document pairs for automatic document summarization. To satisfy the needs of the annotation scheme presented above, the procedure of Daumé III and Marcu was slightly adapted. After the conversion of all phrase-to-phrase alignments into word-to-word alignments by linking each word of the source phrase to each word of the target phrase (all-pairs heuristic), each possible word combination of a given source and target sentence was placed into a specific category, depending on the type of connection between the source and target word.

One-to-one alignments were categorized as *direct links*, whereas words connected within phrase alignments were categorized as *indirect links*. To account for null links, one extra virtual *null word* was added in each source and target sentence, and null links were treated as one-to-null or as null-to-one links. The distinction between regular links and fuzzy links was retained, but regular links within fuzzy links were ignored. This resulted in six different categories: not linked, direct regular links, indirect regular links, direct fuzzy link, indirect fuzzy links and null links.

Kappa was computed over all six categories and results between 0.7 and 0.9 were obtained. Detailed results are presented in table 2. According to Carletta (1996), a Kappa score over 0.8 reflects good agreement, and Kappa values between 0.67 and 0.8 allow tentative conclusions to be drawn. However, as Di Eugenio and Glass (2004) pointed out, it is not easy to compare Kappa scores amongst different annotation tasks as the resulting data sets can ex-

hibit very different characterstics. We therefore do not only rely on Kappa scores, but also calculated *Word Alignment Agreement*, which has been used before to assess inter-annotator agreement for word alignment.

6.2. Word Alignment Agreement

To be able to compare the obtained inter-annotator results with other alignment projects, the Word Alignment Agreement score (Davis, 2002) was calculated. As for Kappa, phrasal alignments were converted into word-to-word alignments using the all-pairs heuristic.

Inter-annotator agreement was measured in terms of similarity between sets of corresponding words. To normalize the interlinked word-to-word links from the phrasal alignments, a weight was assigned to each word-to-word link. The WAA-score is based on the principle that the number of words of the source and target sentence defines the total weight of the alignments of a sentence. For the WAA-score every word contributes 0.5 to the total weight. The total weight of an aligned source and target sentence is hence equal to the number of source words plus the number of target words divided by two.

The WAA score was computed according to the following equation:

$$WAA = \frac{Weight_{Agree}}{Weight_{Total}} \quad (1)$$

The WAA-score is a symmetric measure and gives a number between zero and one, with zero being no agreement and one being perfect agreement. In the inter-annotator experiment WAA-scores between 0.84 and 0.94 were obtained. These results are similar to the scores reported by Melamed (2001b)[4]. Detailed results for all test files are presented in table 2.

	AGREEMENT	
Text	Kappa	WAA
T1	0.73	0.86
T2	0.80	0.86
T3	0.90	0.94
T4	0.71	0.86
T5	0.83	0.92
T6	0.79	0.89
T7	0.73	0.84
T8	0.73	0.94

Table 2: Overview of inter-annotator alignment scores: Kappa score and WAA score.

6.3. Percentage of regular, fuzzy and null links

Apart from the Kappa score and WAA score, we also present the percentage of regular, fuzzy and null links used by each annotator. This overview gives an indication of how *free* or *literal* the translation is, and hence gives an indication of how difficult it was to annotate the text. If the number of words connected by a regular link is very high, the translator stayed close to the source text. A high percentage of fuzzy and null links suggests that the translator took more freedom in translating the text.

	REGULAR		FUZZY		NULL	
Text	Ann1	Ann2	Ann1	Ann2	Ann1	Ann2
T1	88.4	87.2	6.9	7.6	4.8	5.2
T2	87.9	89.8	6.7	4.3	5.4	5.8
T3	93.1	90.8	3.2	4.8	3.7	4.5
T4	85.1	88.0	8.9	5.9	6.0	6.2
T5	94.7	93.2	2.3	2.7	3.0	4.1
T6	91.7	93.3	5.0	3.6	3.3	3.1
T7	90.6	89.2	7.5	6.8	1.9	4.1
T8	94.0	93.6	4.4	3.7	1.7	2.7

Table 3: Percentage of regular, fuzzy and null links used by each annotator.

The inter-annotator agreement rates indicate that the annotators linked the same units most of the time. As expected, most disagreement was found on fuzzy links and null links. Intuitively, paraphrases of complete sentences are the most difficult sentences to align, and annotators often follow a different strategy to link such sentences.

However, the inter-annotator scores seemed sufficiently high to apply the annotation procedure with minor adaptations on the Gold Standard corpus.

7. Corpus

The Gold Standard data set contains texts from different text types. The reference corpus consists of journalistic texts, newsletters and medical European Public Assessment Reports. We assume that for each of the three text types another translation style was adopted, with the journalistic texts being the most free translations and the medical texts being the most literal translations. Table 4 summarizes the formal characteristics of the corpus: total number of words, average sentence length of source and target sentences and the ratio of source-target sentences. In total, the Gold Standard contains more than 25,000 words.

Text type	Total Words	Sentence length (source)	Sentence length (target)	Ratio S/T sentences
Journalistic	7,706	22.0	20.0	0.88
Newsletters	10,480	15.0	15.4	0.99
EPARs	7,536	17.2	17.7	1.01

Table 4: Corpus characteristics of the Dutch-English Gold Standard

7.1. Journalistic texts

The journalistic articles were originally published in The Independent and translated into Dutch for De Morgen, a Flemish quality newspaper. The Independent/De Morgen section in the Dutch Parallel Corpus contains approximately 300,000 words. From this subcorpus three articles were selected for manual annotation. The English source sentences are relatively long, with an average sentence length of 22 words. The ratio source/target sentences

[4]The WAA-score is a further refinement of the metrics used by Melamed.

is 0.88, which means that quite some source sentences are translated by two or more target sentences. This is also reflected in the lower average sentence length of the Dutch target texts (20 words vs. 22 words). The selected set of news articles are characterized by a large percentage of sentences in the source and target texts that do not correspond. It is a local phenomenon that occurs at the beginning and end section of each article.

7.2. Newsletters

The newsletters consist of a collection of newsletters from ING, a Dutch financial institution with diverse international activities. The newsletters bring financial news to private investors. The texts were originally written in Dutch and translated into English. The ING section in the Dutch Parallel Corpus contains more than 180,000 words. From this subcorpus two articles were selected for manual annotation. The average sentence length of the newsletters is relatively short (15 words), but this is mainly due to the structure of the texts: the newsletters consists of short paragraphs, each preceded by a short header.

7.3. European Public Assessment Reports

The EPARs that are included in the Dutch Parallel Corpus originate from one pharmaceutical company. The texts are rather technical with a clear, repetitive structure. The texts were translated from English into Dutch. The EPARs section of the Dutch Parallel Corpus contains more than 600,000 words. From this subcorpus, four EPARs were selected. The average sentence length of the EPARs is 17 words; the ratio source/target sentences is 1.01.

7.4. Overview of links

In table 5 an overview of the different links in the different text types is given. As expected, a different degree of *freeness* can be observed in the different text types, which is reflected in the percentage of fuzzy and null links. The journalistic texts contain the highest number of fuzzy links (11%) and the highest number of null links (9%). The EPARs contain the lowest percentage of fuzzy (6.5%) and null links (3.9%). The newsletters are somewhere in between.

Text type	Regular	Fuzzy	Null
Journalistic texts	79.6	11.1	9.3
Newsletters	88.6	6.9	4.5
EPARs	89.6	6.5	3.9

Table 5: Percentage of regular, fuzzy and null links in the Gold Standard

8. Conclusion

In this paper, we described how we created a Gold Standard for the Dutch-English language pair. In the manual reference corpus three different types of links were used: regular links for straightforward correspondences, fuzzy links for translation-specific shifts of various kinds, and null links for words for which no correspondence could be indicated. As a starting point, annotation guidelines from other word alignment projects were used. To make the manual annotations as useful as possible for different types of projects, a multi-level annotation was introduced in the case of divergent translations: fuzzy links are used to connect paraphrased sections, regular links are used to connect corresponding words within the paraphrased sections.

In order to cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles, the Gold Standard data set contains texts from different text types. We demonstrated that the different writing and translation style was reflected in the number of regular, fuzzy and null links.

The Gold Standard will be publicly available as part of the Dutch Parallel Corpus.

9. Acknowledgements

The DPC project has been carried out within the STEVIN program, which is funded by the Dutch and Flemish Governments.

10. References

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 2002. A system for incremental and interactive word linking. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 485–490, Las Palmas, Spain.

Matthijs Bakker, Cees Koster, and Kitty Van Leuven-Zwart. 2008. Shifts. In Mona Baker and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, pages 269–274. Routledge, London, New York.

Jean Carletta. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(1):249–254.

Hal Daumé III and Daniel Marcu. 2005. Induction of word and phrase alignments for automatic document summarization. *Computational Linguistics*, 31(4):505–530.

Paul C. Davis. 2002. *Stone Soup Translation: The Linked Automata Model*. Ph.d., Ohio State University.

Orphée De Clercq and Maribel Montero Perez. 2010. Data Collection and IPR in Multilingual Parallel Corpora. Dutch Parallel Corpus. In *Proceedings of the Seventh International Conference on Linguistic Resources and Evaluation (LREC-2010)*, Valletta, Malta.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

Patrik Lambert, Adrià De Gispert, Rafael Banchs, and José B Mariño. 2005. Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39:267–285.

Lieve Macken, Julia Trushkina, and Lidia Rura. 2007. Dutch Parallel Corpus: MT corpus and translator's aid. In Bente Maegaard, editor, *Machine Translation Summit XI*, pages 313–320, Copenhagen, Denmark. European Associaton for Machine Translation.

Lieve Macken. 2009. In search of the recurrent units of translation. In Walter Daelemans and Véronique Hoste, editors, *Evaluation of Translation Technology. LANS*

8/2009, pages 195–212. Evaluation of Translation Technology. LANS 8/2009, Brussels, Belgium.

Lieve Macken. 2010. Annotation Guidelines for Dutch-English Word Alignment. Version 1.0. Technical report, Language and Translation Technology Team, Faculty of Translation Studies, University College Ghent.

Dan I. Melamed. 1998. Empirical methods for MT lexicon development. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, Langhorne PA, USA. Springer-Verlag.

Dan I. Melamed. 2001a. *Empirical methods for exploiting parallel texts*. MIT Press, Cambridge, Massachusetts.

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

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

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Canada.

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

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

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