Universiteit Antwerpen

Faculteit Letteren en Wijsbegeerte

Sub-sentential alignment of translational correspondences

De alignatie van overeenkomstige vertaaleenheden onder het zinsniveau

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de Taal- en letterkunde
aan de Universiteit Antwerpen te verdedigen door
Lieve MACKEN

Promotor: Prof. Dr. W. Daelemans
Copromotor: Prof. Dr. J. Buysschaert

Antwerpen, 2010

Print: Silhouet, Maldegem

2010 Uitgeverij UPA University Press Antwerp
UPA is an imprint of ASP nv (Academic and Scientific Publishers nv)
Ravensteingalerij 28
B-1000 Brussels
Tel. + 32 (0)2 289 26 50
Fax + 32 (0)2 289 26 59
E-mail: info@aspeditions.be
www.upa-editions.be

ISBN 978 90 5487 741 7
NUR 616
Legal Deposit D/2010/11.161/065

All translation is a compromise
the effort to be literal
and the effort to be idiomatic.
– Benjamin Jowett

Abstract

The focus of this thesis is sub-sentential alignment, i.e. the automatic alignment of translational correspondences below sentence level. The system that we developed takes as its input sentence-aligned parallel texts and aligns translational correspondences at the sub-sentential level, which can be words, word groups or chunks. The research described in this thesis aims to be of value to the developers of computer-assisted translation tools and to human translators in general.

Two important aspects of this research are its focus on different text types and its focus on precision. In order to cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles, we used parallel texts of different text types. As the intended users are ultimately human translators, our explicit aim was to develop a model that aligns segments with a very high precision.

This thesis consists of three major parts. The first part is introductory and focuses on the manual annotation, the resources used and the evaluation methodology. The second part forms the main contribution of this thesis and describes the sub-sentential alignment system that was developed. In the third part, two different applications are discussed.

Although the global architecture of our sub-sentential alignment module is language-independent, the main focus is on the English-Dutch language pair. At the beginning of the research project, a Gold Standard was created. The manual reference corpus contains three different types of links: 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. The different writing and translation styles in the different text types was reflected in the number of regular, fuzzy and null links.

The sub-sentential alignment system is conceived as a cascaded model consisting of two phases. In the first phase, anchor chunks are linked on the basis of lexical correspondences and syntactic similarity. In the second phase, we use a bootstrapping approach to extract language-pair specific translation patterns. The alignment system is chunk-driven and requires only shallow linguistic processing tools for the source and the target languages, i.e. part-of-speech taggers and chunkers.

To generate the lexical correspondences, we experimented with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries. In the bootstrapping experiments, we started from the precise GIZA++ intersected word alignments. The proposed system improves the recall of the intersected GIZA++ word alignments without sacrificing precision, which makes the resulting alignments more useful for incorporation in CAT-tools or bilingual terminology extraction tools. Moreover, the system's ability to align discontiguous chunks makes the system useful for languages containing split verbal constructions and phrasal verbs.

In the last part of this thesis, we demonstrate the usefulness of the sub-sentential alignment module in two different applications. First, we used the sub-sentential alignment module to guide bilingual terminology extraction on three different language pairs, viz. French-English, French-Italian and French-Dutch. Second, we compare the performance of our alignment system with a commercial sub-sentential translation memory system.

Samenvatting

Het onderwerp van dit proefschrift is het automatisch aligneren van vertaaleenheden onder het zinsniveau. Het systeem dat ontwikkeld werd, vertrekt van zinsgealigneerde parallelle teksten en aligneert vertaaleenheden onder het zinsniveau, namelijk het niveau van woorden, woordgroepen en constituenten (*chunks*). Het onderzoek waarvan in dit proefschrift verslag wordt gedaan, wil van nut zijn voor de ontwikkelaars van computerondersteunde vertaalhulpmiddelen en voor vertalers.

Twee eigenschappen kenmerken dit hele onderzoek. Enerzijds wordt er systematisch gebruik gemaakt van parallelle teksten die tot verschillende tekstsoorten behoren. Dit om een breed scala aan syntactische en stilistische fenomenen af te dekken die voortvloeien uit verschillende schrijfstijlen en vertaalstrategieën. Anderzijds focust dit onderzoek op nauwkeurigheid. Net omdat de eindgebruikers die we voor ogen hebben menselijke vertalers zijn, was het onze doelstelling om een systeem te ontwikkelen dat eenheden kan aligneren met een heel hoge graad van precisie.

Het proefschrift bestaat uit drie grote delen. Het eerste daarvan is inleidend en beschrijft het manueel aangemaakte referentiecorpus, de basiscomponenten (parallelle corpora en NLP-software) waarvan het systeem gebruik maakt en de evaluatiemethoden. De belangrijkste bijdrage van dit proefschrift – de beschrijving van het ontwikkelde alignatiesysteem – vormt het tweede deel. In het derde deel worden twee toepassingen besproken.

Hoewel de globale architectuur van onze alignatiemodule taalonafhankelijk is, ligt de nadruk in dit onderzoek op de talencombinatie Engels-Nederlands. Bij de aanvang van het project werd er een referentiecorpus aangemaakt waarin manueel de vertaaleenheden onder het zinsniveau werden gealigneerd. Drie soorten links werden onderscheiden: *reguliere* links voor ongecompliceerde correspondenties, vage of *fuzzy* links voor verschuivingen van diverse aard, en zeroverbindingen of *null links* voor elementen die werden toegevoegd of weggelaten tijdens de vertaling.

De alignatiemodule is opgebouwd volgens een cascade-model bestaande uit twee fasen. In de eerste fase worden ankerconstituenten (*anchor chunks*) met elkaar verbonden op basis van lexicale correspondenties en syntactische overeenkomsten. In de tweede fase gebruiken we een *bootstrapping* benadering om taalpaarspecifieke vertaalpatronen te extraheren. Het alignatiesysteem maakt gebruik van woordalignaties en oppervlakkige taalkundige analyses van de bron- en de doeltekst, d.w.z. woordsoorten en constituentinformatie.

Voor het genereren van de lexicale alignaties experimenteerden we met twee soorten bilinguale woordenboeken: de ene soort handmatig opgesteld, de andere soort probabilistisch. In de bootstrapping experimenten vertrekt het systeem van de woordalignaties die gegenereerd worden door GIZA++. Het door ons ontwikkelde systeem is in staat meer alignaties te genereren dan GIZA++ zonder op precisie in te boeten. Bovendien kan het systeem niet-aaneensluitende eenheden aligneren wat noodzakelijk is voor talen waarin gescheiden werkwoordsgroepen en scheidbare werkwoorden vaak voorkomen.

In het laatste deel van dit proefschrift bespreken we twee verschillende toepassingen. Enerzijds gebruiken we een aangepaste versie van onze alignatiemodule voor bilinguale terminologie-extractie uit drie verschillende talencombinaties, namelijk Frans-Engels, Frans-Italiaans en Frans-Nederlands. Anderzijds vergelijken we de uitvoer van ons alignatiesysteem met dat van een commercieel beschikbaar vertaalgeheugen dat werkt met eenheden onder het zinsniveau.

Acknowledgements

What started out as an adventure – pursuing a PhD was part of the job description after all – turned out to be a fascinating although high-demanding activity. Having been a team leader in a high-tech company in a previous life, I viewed my PhD research as a one-person project with the luxury of having no one else to blame but myself in the case of failure. Nevertheless this work would not have been possible without the help and support of many people.

I would like to take this opportunity to express my great appreciation and gratitude to my supervisors Walter Daelemans and Joost Buysschaert. They both gave me considerable freedom. It was Joost who introduced me to the wonderful life of computer-aided translation. Living in Ostend, working in Ghent and having a supervisor in Antwerp inevitably led to a long-distance relationship with Walter, which turned out to work pretty well. It forced me to prepare my meetings with Walter thoroughly and because Walter looked at the issues from a different angle, I often left Antwerp with fresh ideas and new research paths to be explored.

I also wish to thank Rita Godyns and Sonia Vandepitte. First of all for hiring me, but also for their belief in the benefit of language technology for the Faculty of Translation Studies. Also a special word of thanks to Willy Vandeweghe for his trust in me, which allowed us to start the DPC adventure together.

The number of colleagues grew every year and the small research unit dealing with language technology developed into a full-fledged research group under the

leadership of Véronique Hoste. I want to thank Véronique for creating this stimulating research environment.

Also special thanks to Els Lefever for being my room mate for so many years. But also for her humour and her ability to put things into perspective, which made the tedious process of manual annotation and the endless cycles of preprocessing that are part of any NLP research project easier to bear.

I am especially grateful to Orphée Declercq who had the energy to proofread my final text. Also thanks to all other LT3 team members: Bart, Philip, Peter, Klaar, Isabelle, Dries, Kathelijne, Nele, Sofie and Marianne, for configuring the server, for installing Moses, for taking over my teaching duties so I could concentrate on writing this thesis, for offering lodging in Ghent, for getting sandwiches, for the lunch breaks, and for all the birthday cakes.

But most words of thanks are reserved for my family. A warm thanks goes to my children Annelies, Dries and Matthias for being patient, understanding, and proud, and for constantly reminding me of the things that are really important in life. Finally, I would like to acknowledge my husband Thomas. Without his daily support this work would never have been completed. Thomas combined his own job with the daily care of the children, and prepared me a meal every evening I commuted from Ghent. Thanks Thomas.

Ghent, March 2010

Contents

CHAPTER 1

Introduction

It is widely acknowledged that parallel texts, i.e. original source texts and their translations are a useful resource for professional translators to solve translation difficulties. Pierre Isabelle stated already in 1993 that “existing translations contain more solutions to more translation problems than any other available resource” (Isabelle et al. 1993).

The re-use of previous translations by human translators is also the concept behind translation memory systems, in which aligned source and target sentences are stored in a database. During translation, the translation memory system matches the new source sentence with the source sentences in its database and proposes previously translated sentences to the translator. The system can either return sentence pairs with identical source segments (exact matches) or sentences that are similar, but not identical to the sentence to be translated (fuzzy matches). As the operational unit of the first generation translation memory systems is the sentence, such systems are only useful for text types where nearly full-sentence repetition frequently occurs: technical documents, texts with related content or text revisions. To go beyond this limitation, second generation translation memory systems provide additional translation suggestions for sub-sentential chunks.

Bowker and Barlow also point out the usefulness of sub-sentential units in their paper on translation memories and bilingual concordancing systems:

> Nevertheless, there is still a level of linguistic repetition that falls between full sentences and specialized terms – repetition at the level of expression or phrase. This is in fact the level where linguistic repetition will occur most often. (Bowker and Barlow 2004)

In the domain of statistical machine translation, in which all the knowledge needed for translation is extracted from parallel texts, the currently best performing systems are based on phrase-based models (Koehn 2009), which in fact assemble translations of different sub-sentential units. The sub-sentential units are sometimes defined as contiguous sequences of words; in other cases more linguistically motivated definitions are used.

This thesis is about sub-sentential alignment, i.e. the automatic alignment of translational correspondences below the level of the sentence. The research described in this thesis is situated in the field of language technology, and more specifically translation technology. It aims to be of value to the developers of computer-assisted translation tools and to human translators in general.

The system that we have developed takes as its input sentence-aligned parallel texts and aligns translational correspondences at the sub-sentential level, which can be the level of words, word groups or chunks. Our research can thus be situated on the interface between sentence alignment and word alignment, two well-studied areas of research.

We conceive our sub-sentential alignment system as a cascaded model consisting of two phases. In the first phase, anchor chunks are linked on the basis of lexical correspondences and syntactic similarity. In the second phase, we use a bootstrapping approach to extract language-pair specific translation patterns. The alignment system is chunk-driven and requires only shallow linguistic processing tools for the source and the target languages, i.e. part-of-speech taggers, lemmatizers and chunkers.

To generate the lexical correspondences, we experimented with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries. In the bootstrapping experiments, we started from the precise GIZA++ intersected word alignments. The proposed system improves the recall of the intersected GIZA++ word alignments without sacrificing precision, which makes the resulting alignments more useful for incorporation in CAT-tools or bilingual terminology extraction tools. Moreover, the system's ability to align discontiguous chunks makes the system useful for languages containing split verbal constructions.

In this thesis, we use parallel texts of different text types to cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles. At the start of our research project only a few text types were available. At a later stage and thanks to our involvement in the compilation of the Dutch Parallel Corpus Project (Macken, Trushkina and Rura 2007), a more diverse parallel corpus came at our disposal.

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. The DPC project was a collaborative project between the University of Leuven Campus Kortrijk and the Faculty of Translation Studies of University College Ghent. As partner of the core research team, we were closely involved in all processing steps of the DPC project.

Another important aspect of this thesis is its focus on precision. By design, already at the outset of our project, the intended users were human translators. The central research question of this project was to what extent it was possible to develop a robust automatic sub-sentential alignment module that is useful for human translators. As erroneous translation suggestions can be more distracting than beneficial to translators, our explicit aim was to develop a model that aligns segments with a very high precision.

Although the overall architecture of our sub-sentential alignment module is language-independent, the main focus is on the English-Dutch language pair. In chapter 7, however, we demonstrate how the sub-sentential alignment module was used to guide bilingual terminology extraction for French-English, French-Italian and French-Dutch.

In this introductory chapter, we briefly discuss relevant literature within the domain of translation studies. At the end of this chapter an outline of the thesis is presented.

1.1 Insights from translation studies

At least two areas in the field of translation studies are of relevance to this thesis. The first one is the product-oriented approach, which has a long-standing tradition in translation studies, and whose focus is on the translation product. The second one is the process-oriented approach, which attempts to gain access to the translator's mind during the translation process.

1.1.1 Product-oriented approach

When comparing source texts and their translations, translational correspondences are often difficult to determine at the word level as word-by-word correspondences can only be found for a limited number of words in an average sentence pair. The remaining correspondences are at the level of combination of words. It is therefore not surprising that there is a vast literature on equivalence, translation shifts and translation strategies and over the years different theories have been proposed.

As early as 1958, Vinay and Darbelnet developed a detailed taxonomy of translation procedures. They defined their unit of translation as "the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually" (Hatim and Munday 2004, p. 18). But it was Catford (1965) who introduced the term *shift in translation*. The term *shifts* commonly refers to changes which occur or may occur in the process of translating (Bakker, Koster and Van Leuven-Zwart 2008).

Catford distinguishes two major types of shifts: *level* shifts (e.g. shifts from grammar to lexis in distant languages) and *category* shifts (e.g. changes in word order or word class as in the following English-Dutch translation *to treat* ∼ *voor de behandeling van* [En: *for the treatment of*], in which the verb *treat* is translated by a noun *behandeling*). Another distinction that is often made is that between *obligatory* and *optional* shifts: the former refer to shifts that are imposed as a result of differences in the language systems, whereas the latter term is used to indicate optional choices of the translator.

Depending on the aspect of translation under investigation, the units the theories operate on differ from words and word groups (Vinay and Darbelnet 1958, Catford 1965) to sentences and even texts (Newmark 1988), and the phenomena they describe are situated on the syntactic, semantic and pragmatic level.

Van Leuven-Zwart (1989, 1990) was the first to make a detailed attempt at producing and applying a model of shift analysis in literary texts. She divided the source and target texts into comprehensible textual units called *transemes* and built up a classification of all the shifts at the microstructural level. In total, a taxonomy of 37 subcategories of translation shifts was introduced. The model of Van Leuven-Zwart has hardly been adopted by other researchers. Perhaps, with 37 subcategories of translation shifts, the model may have been too complex.

More recent attempts were undertaken by Thunes (1998), Merkel (1999b) and Cyrus (2006). Thunes limits her analysis to finite clauses from texts of different domains and classifies them according to varying degrees of translational complexity. In his unpublished PhD thesis, Merkel provides a very detailed analysis of 100 sentence pairs in different text types. Although he did not attempt to automate the model, he points out the possibility that some of the tags describing structural and syntactic shifts could be generated automatically (Merkel 1999b, pp. 208–209). Cyrus (2006) annotated grammatical and semantic shifts on the basis of predicate-argument structures in English-German texts taken from Europarl (Koehn 2005).

The literature on translation shifts suggests that the changes occurring in the process of translating can be classified according to certain regularities. An analysis of translation shifts in general and semantic and pragmatic shifts in particular falls outside the scope of this thesis. However, in the work presented here, we extract frequently occurring translation patterns – mainly at the structural level – and use them to align more complex translational correspondences automatically. Some English-Dutch examples are given below:

- DET+ADJ+N → ADJ as in *a fragile thing* → *kwetsbaar.*
- PREP+V-prpa → PREP+DET+N+PREP as in *at inventing* → *in het verzinnen van.*

1.1.2 Process-oriented approach

The process-oriented view aims to gain insight in the mental processes of translation. Among the research questions in this approach, we mention some that are of interest to us:

- On what kind of text units do translators operate?
- Do these units correspond with syntactic units (e.g. constituents, clauses, sentences)?
- How do translators solve translation problems?

In the 1980s think-aloud protocols (TAPs) were introduced in translation studies to investigate the processes that take place when translating a text. The methods are adopted from cognitive psychology and involve asking translators to verbalize their thoughts while translating a text. The verbalizations are recorded and analyzed. The theoretical framework for such experiments is described in Ericsson and Simon (1993/1984). An overview of think-aloud protocols in translation research is given in Bernardini (2001) and an annotated bibliography is provided by Jääskeläinen (2002).

Lörscher (1992, 2005) collected translations of both advanced foreign language learners and professional translators together with concurrent verbalizations. The transcripts were analyzed and a number of translation strategies – procedures which the subjects employ in order to solve translation problems – were identified. The results showed that the length of the translation units was larger with professional translators than with foreign language students:

> The units of translation, i.e. the SL text segments which subjects extract and put into their focus of attention in order to render them into the TL *as a whole*, are considerably larger among professional translators than among foreign language students. In other words, the processing system of professionals can obviously address larger units than that of non-professionals. The former mainly choose phrases, clauses or sentences as units of translation whereas the latter concentrate on syntagmas and especially on single words. (Lörscher 2005, p. 605)

This difference between foreign language learners and professionals is highly intuitive. However, for Bernardini, this distinction is not so clear-cut:

> According to most researchers, the length of translation units is an indication of proficiency [...] Clearly, this does not mean that a professional translator never stops midway through a sentence, but only that the sentence is processed as a unit with more local problems tackled on the way. The suggestion can be put forward that attention units are better defined in hierarchical rather than sequential terms, with smaller units being processed within larger units. (Bernardini 2001, p. 249)

One of the criticisms on think-aloud protocols is that the think-aloud condition has not only an effect on the speed of the translation task, but also interferes with the translation process itself. Jakobsen (2005) observed that the think-aloud condition slowed down the translation process with 25% and forced the translators to work with smaller segments.

Another - complementary - method of digging into the process of translation is keystroke logging. During a translation task, keystroke-recording software logs all keystrokes – including backspaces and deletions – and records the time the strokes were made. It can thus provide rich data about the translation processes as it can reveal draft solutions, false starts, revisions, and the like. The pause information between the strokes gives a clue to the segments the translators worked with. In practice, keystroke logging and thinking-aloud are often combined in experiments. Two studies of this type are of particular interest to us.

The first one is a study by Asadi and Séguinot (2005) in which the translation strategies of professional translators in their normal working environment were analyzed. They observed two different production styles. The first one is a more holistic approach to translation, what they called the *prospective thinking* style, which was characterized by the processing of large proposition- or sentence-length segments and translating them mentally before typing the solution. The second one was called *on-screen translation* and was characterized by the processing of shorter phrase-like segments. The translators belonging to the second group translated as they read the text, followed the source language syntax and lexical items closely until a more idiomatic solution was recognized. In the revision phase, the sentences were further adapted to fit the target language more appropriately.

The study by Ronowicz et al. (2005) on the use of dictionaries in the translation process reports similar trends: some professional translators read the whole text before they started translating while others started the translation process after having read a paragraph; the translators were very cautious about the naturalness of word usage and tried out several equivalents; the translators reworked their target texts several times until they were satisfied with the target text.

Yet another type of study examined the log files of *TransSearch*, a bilingual concordancing system which is widely used by Canadian translators (Simard and Macklovitch 2005). They examined the queries formulated by the translators: 85% of the queries consist of two or more contiguous words; most of the queries are related to one or more syntactic chunks or several syntactic chunks.

The latest methods in process-oriented translation studies combine keystroke logging and eye-tracking data. Eye-tracking data, and more specifically gaze information, provides additional information about the translation process. While keystrokes reveal information about the text production phase, gaze information focuses on the text understanding phase in the translation process. This new methodology is expected to bring new insights on the translation process as a whole. However, this line of research is still in its infancy at the moment of writing, and no descriptions of large-scale experiments can be found in the literature. A first attempt to integrate the gaze and keystroke data in a visualization tool is described in Spakov and Räihä (2008). A small-scale analysis of a translation session can be found in Carl (2009).

The literature review reveals the complexity of the translation process. Segments of different length can be used as a translation unit in the course of translation activity. Several studies point out that the length of translation units correlates with proficiency. This would suggest that sub-sentential translation suggestions are mainly useful for translation learners.

However, the study of Asadi and Séguinot (2005) sheds another light on this issue by distinguishing two types of translators: *prospective thinkers* and *on-screen translators*. The findings of Asadi and Séguinot are encouraging for our research as the on-screen translators form the potential user group of more intelligent computer-aided translation tools that can offer several translation alternatives for specific translation problems. According to Simard and Macklovitch (2005) those specific translation problems often consist of more than two words and take the form of syntactic chunks.

The sub-sentential alignment module that we have developed is primarily chunk-driven. We expect that the aligned chunks can assist human translators in their search for the most appropriate translation equivalent.

1.2 Thesis outline

This thesis contains nine chapters presenting research that has been carried out over the last six years. Most chapters contain a more detailed version of work that has been presented at conferences and published in conference proceedings. Consequently, this research benefited from the comments of many anonymous reviewers and discussions at conferences. References to the publications are included in the text.

Chapter 2 focuses on the resources that were used. It discusses available parallel corpora and techniques to develop such corpora. In this chapter more details are given on the compilation of the Dutch Parallel Corpus. It discusses sentence alignment and word alignment techniques and describes the handcrafted and the probabilistic bilingual dictionaries that were used in the experiments. It describes the tools that were used to provide additional linguistic annotations: lemmata, part-of-speech classes and base chunks.

Chapter 3 reports on the creation of the manually annotated corpora that will be used throughout this thesis as test and development corpora. It introduces the annotation scheme, the annotation guidelines and the annotation environment, describes the inter-annotator experiment that has been carried out and demonstrates the importance of including different text types.

Chapter 4 deals with evaluation metrics. The evaluation of word alignment or chunk alignment systems is not a trivial task. Different evaluation metrics exist, and they mainly differ in the way divergent translational correspondences are treated. In this chapter we discuss and compare the different evaluation metrics that will be used throughout this thesis.

Chapters 5 and 6 represent the core part of this thesis. Chapter 5 gives an overview of the general architecture of our system and presents in detail the *Anchor Chunk Alignment* step. We report on experiments with two different types of bilingual dictionaries to generate the lexical correspondences: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries of various sizes. Chapter 6 focuses on the alignment of more complex translational correspondences. In this chapter, we present a chunk-driven bootstrapping approach to extract translation patterns and compare the performance of our system with state-of-the-art methods used in phrase-based statistical machine translation.

The last three chapters focus on applications. In Chapter 7, we describe how the sub-sentential alignment module was used to guide bilingual terminology extraction. In this chapter we also demonstrate the language independence of our approach as the terminology extraction experiments were carried out on three different language pairs, viz. French-English, French-Italian and French-Dutch.

Chapter 8 compares the performance of a sentence-based translation memory system with a sub-sentential translation memory system on different text types. We evaluate the chunk-alignment module of a sub-sentential translation system and compare its performance with our system.

Chapter 9 concludes this thesis with a summary of contributions and prospects for future research.

CHAPTER 2

Resources

2.1 Introduction

This chapter focuses on the resources that were used in this thesis. The subsentential alignment system that has been developed takes as its input sentence-aligned parallel texts that are enriched with different types of linguistic analysis (part-of-speech codes, lemmata and chunk boundaries). The system also relies on a bilingual dictionary to retrieve lexical correspondences. The work presented here, therefore builds on research on parallel corpora, sentence and word alignment techniques and linguistic analysis.

In this chapter, we describe available parallel corpora for the English-Dutch language pair and techniques to develop such corpora. We provide more details on the sentence alignment process adopted in the Dutch Parallel Corpus project. We discuss word alignment techniques and describe the handcrafted and probabilistic bilingual dictionaries that were used in the experiments. We end this chapter with a description of the tools that were used to perform a shallow linguistic analysis.

2.2 Parallel corpora

Parallel corpora are an indispensable resource for this dissertation. By definition, parallel corpora contain texts – in different languages – that are translations of each other. The translations can be *direct*, i.e. original source language texts that were translated into a specific target language, or *indirect*, i.e. original source language texts that were translated via an intermediate language into a specific target language. Parallel corpora can contain texts originating from one specific text provider (e.g. the proceedings of the European Parliament plenary debates), or can contain texts belonging to different text types[1] (e.g. the Dutch Parallel Corpus (see section 2.2.3).

Even if the aim is the creation of a balanced parallel corpus, the data acquisition process is determined by a number of limitations such as the availability of translated data, the quality of translated material and the proportional availability of translated material for all targeted languages and translation directions (Olohan 2004, p. 25).

In this dissertation, data from three different sources are used: Europarl, eCoLoRe, and DPC.

2.2.1 Europarl

The Europarl corpus[2] (Koehn 2005) contains the proceedings of the European Parliament plenary debates in each official language of the European Union. The texts are a written reproduction of prepared speeches and cover a wide range of subject fields. Due to its language coverage and size (20 to 30 million words per language), it is the most widely used corpus in the domain of statistical machine translation. We made use of the second version of the Europarl corpus, which contains data from April 1996 to September 2003.

Not all translations of Europarl suit our purposes. Most statistical machine translation systems do not take into account translation direction, but rather make use of all available parallel texts. Only recently has the awareness grown in the MT research community that translated text differs from non-translated text and that taking into account directionality when training a statistical machine translation system might have an impact on MT quality (Kurokawa, Goutte and Isabelle 2009).

[1] *Text types* differ from *genres* in that they are based on text-internal linguistic criteria (Lee 2001, p. 38).

[2] http://www.statmt.org/europarl

The situation in the Europarl corpus is very complex, as English is used as intermediate language in translation:

> All members of the European Parliament have the right to use their native language in the plenary sessions of parliament. This does not mean that they always use that right. The process of translating the plenary debates with 20 working languages would be unmanageable if the full reports of the meetings were only translated directly. Therefore, English is used as pivot language (intermediate language) for the translation of the full reports from all languages to all languages (De Groot 2006)[3].

This means that for all language pairs without English as source or target language only indirect translations are available. As we preferred to work with *direct* translations of original English source sentences produced by native English speakers into Dutch, we made use of the language tag and speaker identity tag, which are available for every intervention. The language tags allowed us to extract all direct translations from English into Dutch. The speaker identity tags allowed us to extract all English source sentences produced by native English speakers by matching them against a compiled list of British and Irish Members of Parliament. The result was an English-Dutch subcorpus of 6 million words.

2.2.2 eCoLoRe

The eCoLoRe project[4] (2002-2005) provides resources to support computer-assisted translation (CAT) training. The resources include guidelines and training kits covering different text types, file formats and language pairs. Apart

[3] Original Dutch citation: "Alle leden van het Europees Parlement hebben het recht zich in de plenaire vergaderingen van het EP van hun moedertaal te bedienen. Dit wil niet zeggen dat zij van dat recht ook altijd gebruik maken. De vertaal- en vertolkingslogistiek van de plenaire debatten (officieel "Volledig verslag van de vergaderingen" genoemd) zou met 20 werktalen onbeheersbaar zijn als uitsluitend rechtstreeks zou worden vertaald. Daarom is gekozen voor het gebruik van Engels als spiltaal (tussentaal) voor de vertaling van het volledig verslag uit alle talen naar alle talen. De procedure is als volgt: De op band opgenomen interventies (in 20 talen) worden eerst uitgetikt en vervolgens "gefatsoeneerd" door een vertaler met de betreffende taal als moedertaal. De aldus uit 20 verschillende talen samengestelde en gefatsoeneerde lappendeken wordt, voor alle andere talen dan het Engels, vervolgens in het Engels vertaald en vanuit die taal naar de 19 overige talen. Uiteraard blijven voor, bijvoorbeeld, de NL-versie van het volledig verslag de oorspronkelijk in het Nederlands uitgesproken teksten onveranderd (behalve dan het fatsoeneren)."

[4] http://ecolore.leeds.ac.uk

from the texts to be translated, the training kits also include translation memories, i.e. the source and target texts. In this dissertation, we made use of the translation memories of the English-Dutch SAP user manuals.

2.2.3 Dutch Parallel Corpus

Since high-quality parallel corpora with Dutch as a central language did not exist or were not accessible for the research community due to copyright restrictions, the compilation of aligned parallel corpora with Dutch as a central language was one of the priorities of the STEVIN program (Odijk et al. 2004).

The Dutch Parallel Corpus project (2006-2009) aimed at fulfilling this need. In the DPC 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, namely administrative texts, texts dealing with external communication, literary texts, journalistic texts and instructive texts, and is balanced with respect to text type and translation direction. A part of the corpus is trilingual and contains Dutch texts translated into both English and French. The English-Dutch part of the corpus consists of 2.5 million words.

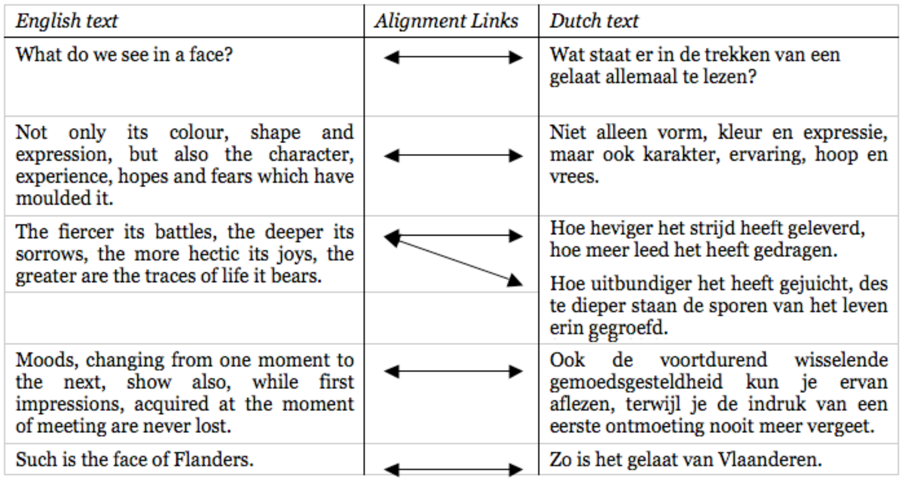
The DPC project was a collaborative project between the University of Leuven Campus Kortrijk and the Faculty of Translation Studies of University College Ghent. As partner of the core research team, we were closely involved in all processing steps of the DPC project. The DPC project has been carried out within the STEVIN program and was funded by the Dutch and Flemish governments. The DPC will be made available to the research community via the HLT-Agency[5].

2.3 Sentence alignment

A first processing step to make parallel corpora useful for further research is sentence alignment. Sentence alignment is the process of finding equivalent text chunks at the level of the sentence in parallel texts. The sentences linked by the alignment procedure represent translations of each other in different languages. An example is given in figure 2.1.

Most sentence alignments are 1:1, 1:many or many:1 alignments. Zero alignments occur when no translation can be found for a sentence of either the

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



English text	*Alignment Links*	*Dutch text*
What do we see in a face?	↔	Wat staat er in de trekken van een gelaat allemaal te lezen?
Not only its colour, shape and expression, but also the character, experience, hopes and fears which have moulded it.	↔	Niet alleen vorm, kleur en expressie, maar ook karakter, ervaring, hoop en vrees.
The fiercer its battles, the deeper its sorrows, the more hectic its joys, the greater are the traces of life it bears.	↔ ↘	Hoe heviger het strijd heeft geleverd, hoe meer leed het heeft gedragen.
		Hoe uitbundiger het heeft gejuicht, des te dieper staan de sporen van het leven erin gegroefd.
Moods, changing from one moment to the next, show also, while first impressions, acquired at the moment of meeting are never lost.	↔	Ook de voortdurend wisselende gemoedsgesteldheid kun je ervan aflezen, terwijl je de indruk van een eerste ontmoeting nooit meer vergeet.
Such is the face of Flanders.	↔	Zo is het gelaat van Vlaanderen.

Figure 2.1: Example of 1:1 and 1:2 sentence alignments (extracted from the DPC).

source or the target language, i.e. when a corresponding part of text is missing in the other language. Many-to-many alignments are used to model overlapping alignments. In most parallel corpora, crossing alignments cannot be modelled and are grouped as many-to-many alignments.

A range of tools and algorithms are available for the task of sentence alignment. The algorithms are mainly based on two different approaches: the sentence-length-based approach and the word-correspondence-based approach.

In the sentence-length-based approach the alignment process is guided by the assumption that the lengths of corresponding sentences are highly correlated. In other words, short sentences tend to be translated by short sentences, and long sentences by long sentences or several short sentences. The sentence-length-based approach was introduced by Gale and Church (1991) and Brown, Lai and Mercer (1991). A probabilistic score is assigned to each proposed correspondence of sentences and the scores are used in a dynamic programming framework to find the maximum likelihood alignment of sentences. Structural information – i.e. headers, titles, paragraph information and the like – can be used to restrict the space of allowable alignments.

The word-correspondence-based approach is described in the seminal paper of Kay and Röscheisen (1993) and is based on the assumption that if sentences are translations of each other, the corresponding words must be translations as well. Their algorithm performs both sentence and word alignment and both processes reinforce each other. While Kay and Röscheisen used only text-internal information and derived the word correspondences from the texts to be aligned, an intuitive extension is the use of electronically available bilingual dictionaries (Melamed 1997). A second extension is the use of cognates – i.e. word tokens that are graphically identical or similar such as proper names, dates, certain symbols and the like – as corresponding words (Simard et al. 2000).

The performance of the individual alignment tools varies for different types of texts and language pairs and normally, a manual verification step is necessary to guarantee high quality of the data. In the framework of the DPC project, we carried out a formal evaluation of three different approaches and showed that by combining the output of different aligners the amount of manual work necessary to achieve near 100% accuracy can be reduced significantly. The three different alignment tools used in the experiments and the evaluation are described below.

The *Vanilla Aligner* (Danielsson and Ridings 1997) is an implementation of the sentence-length-based statistical approach of Gale and Church (1993). As its input, the Vanilla aligner expects texts split into sentences and paragraphs. The numbers of paragraphs in the source and target texts should be equal. The tool assumes that the paragraphs are aligned and finds sentence alignments within the aligned paragraphs.

The *Geometric Mapping and Alignment* (GMA) developed by Melamed (1997) is a hybrid approach, which is based on word correspondences and sentence length. To establish the word correspondences, the system relies on finding cognates and – optionally – translation lexicons. In the DPC project, we used the NL-Translex translation lexicons (see section 2.4.1) as additional source for establishing word correspondences.

The *Microsoft Bilingual Aligner* developed by Moore (2002) uses a three-step hybrid approach to sentence alignment. The aligner uses sentence length and lexical correspondences, both of which are derived automatically from the source and target texts. In a first step, an initial set of high accuracy alignments is established using a sentence-length-based approach. In the second step, this initial set of alignments serves as the basis for training a statistical word alignment model (Brown et al. 1993). Finally, the corpus is realigned, augmenting the initial set of alignments with sentences aligned on the basis of the word alignments. The aligner outputs only 1:1 links and disregards all other alignment types.

A reference alignment was created for approximately 1 million words. The reference corpus consists of two language pairs (Duch-English and Dutch-French) and contains texts belonging to five different text types. More details on the reference corpus are given in table 2.1.

	Dutch-English		Dutch-French	
	Aligned Pairs	Words	Aligned Pairs	Words
Administrative	2,265	78,156	2,297	81,187
External Com.	3,914	140,968	4,373	171,427
Instructive	3,444	87,004	2,380	68,651
Literature	1,458	57,359	1,059	48,911
Journalistic	3,474	131,477	2,851	98,519
Total	14,555	494,964	12,960	468,695

Table 2.1: Size of the different subcorpora expressed in number of aligned sentence pairs and total number of words.

In table 2.2, the results are given for the three different alignment tools separately and for different combinations of the output of the alignment tools. In the combined output, only the alignments on which the different tools agreed were retained. The results are given in terms of precision and recall, which are defined as follows:

$$Precision = \frac{CorrectAlignments}{ProposedAlignments} \tag{2.1}$$

$$Recall = \frac{CorrectAlignments}{ReferenceAlignments} \tag{2.2}$$

The results show that the alignment tools separately obtain precision scores between 91 and 95%. However, if the output of all three alignment tools is combined, 100% precision can be obtained at a cost of reduced recall scores. The recall figures could be improved by retaining all alignments established by at least two alignment tools.

As the aim of the DPC project was the creation a high-quality sentence-aligned parallel corpus, using different alignment tools had some advantages. On the one hand, with a precision of nearly 100%, the links on which at least two alignment tools agreed could be considered of high quality. On the other hand, to guarantee high-quality alignments for the whole corpus, only a small portion of the corpus (viz. the links on which the different tools did not agree) needed manual verification.

Therefore, in order to process the 10 million words of the DPC, we combined the output of the three alignment tools in order to minimize the human effort to correct the data. All alignments that were not established by at least two alignment tools were flagged for manual verification.

	Dutch-English		Dutch-French	
	Prec	Rec	Prec	Rec
Vanilla	0.94	0.93	0.94	0.93
GMA	0.93	0.92	0.91	0.90
MS	0.95	0.87	0.93	0.82
All three aligners	1.00	0.83	1.00	0.77
Vanilla + GMA	0.99	0.89	1.00	0.85
Vanilla + MS	1.00	0.86	1.00	0.81
At least two aligners	0.99	0.91	1.00	0.89

Table 2.2: Precision and recall results for the three alignment tools separately and combined on a Dutch-English and Dutch-French manually verified subcorpus of the DPC.

An error analysis revealed the strengths and weaknesses of the three alignment tools:

- GMA sometimes suffers from an *error recovery problem*. On several occasions, GMA proposed large alignment blocks (e.g. a 1-9 alignment) when the reference contained a zero alignment. As the Vanilla aligner uses paragraph alignment information, the errors of the Vanilla aligner were situated on a more local level.

- GMA performs better than the Vanilla aligner on modeling zero alignments which probably can be attributed to the use of word-correspondence information.

- The Vanilla aligner performs better than GMA on 1:2, 2:1, and n:m alignments, which might be due to the fact that it only uses sentence-length-based information.

- The Microsoft aligner only outputs 1:1 alignments. Therefore, we expected it to be the aligner with the highest precision, which is not the case for the Dutch-French data. As the system needs large (homogeneous) data sets to build a reliable bilingual dictionary, these results can be most probably ascribed to data scarcity.

Tables 2.3 and 2.4 present the results broken down by text type. The results demonstrate that the performance varies across different text types. For both language pairs, the instructive and administrative texts are the easiest to align and the journalistic texts are the most difficult to align. An intuitive explanation is that instructive and administrative texts are typically translated rather literally and therefore contain a higher degree of 1:1 correspondences. Journalistic texts on the other hand demonstrate a more free translation style and contain more occurrences of zero alignments.

	Adm		ExtCom		Instr		Lit		Journ	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Vanilla	0.97	0.95	0.94	0.93	0.97	0.96	0.95	0.93	0.90	0.87
GMA	0.96	0.95	0.94	0.93	0.93	0.92	0.95	0.93	0.90	0.87
MS	0.97	0.91	0.95	0.88	0.96	0.91	0.98	0.89	0.92	0.77
All three	1.00	0.87	1.00	0.84	01.00	0.89	1.00	0.85	0.99	0.74
Van+GMA	0.99	0.91	0.99	0.89	0.99	0.91	0.99	0.90	0.98	0.83
Van+MS	1.00	0.89	1.00	0.87	1.00	0.92	1.00	0.87	0.99	0.75
Min. two	0.99	0.93	0.99	0.92	0.99	0.94	0.99	0.92	0.98	0.85

Table 2.3: Precision and recall results for the three alignment tools separately and combined on Dutch-English subcorpora of different text types.

	Adm		ExtCom		Instr		Lit		Journ	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Vanilla	0.95	0.95	0.95	0.94	0.97	0.97	0.95	0.95	0.88	0.85
GMA	0.95	0.93	0.92	0.91	0.94	0.93	0.94	0.95	0.81	0.80
MS	0.95	0.88	0.94	0.84	0.96	0.91	0.93	0.81	0.85	0.69
All three	1.00	0.84	1.00	0.78	1.00	0.87	1.00	0.77	1.00	0.63
Van+GMA	1.00	0.89	1.00	0.87	1.00	0.91	0.99	0.89	0.99	0.75
Van+MS	1.00	0.88	1.00	0.82	1.00	0.91	1.00	0.80	1.00	0.67
Min. two	1.00	0.93	1.00	0.90	1.00	0.95	0.99	0.92	0.99	0.79

Table 2.4: Precision and recall results for the three alignment tools separately and combined on Dutch-French subcorpora of different text types.

2.4 Bilingual dictionaries

Next to sentence-aligned parallel texts, our sub-sentential alignment system makes use of a bilingual dictionary in order to retrieve lexical correspondences.

We experimented with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries that are derived from a parallel corpus by using a statistical word alignment package. In chapter 6, the system does not start from a bilingual dictionary but builds further on the output of GIZA++, a state-of-the-art tool for statistical word alignment.

2.4.1 Handcrafted bilingual lexicon

The handcrafted bilingual lexicon was derived from the English-Dutch and Dutch-English NL-Translex lexicons[6](Goetschalckx, Cucchiarini and Van Hoorde 2001). The English-Dutch part of NL-Translex contains 104,116 entries; the Dutch-English part 49,697 entries. The two NL-Translex dictionaries contain lemmata.

Apart from one-to-one correspondences (e.g. *force - kracht*) and compounds (e.g. *market sector - marktsegment*), the NL-Translex lexicons also contain a large number of phrasal correspondences (e.g. *come into force - van kracht worden*, *agreement on government procurement - overeenkomst inzake overheidsopdrachten*).

As the major challenge of our research project is the automatic alignment of those more complex phrasal correspondences, we wanted to be able to include these. However – because of the way the data was represented – it was not possible to automatically distinguish between compounds and phrasal correspondences. Therefore, we only retained all one-to-one (i.e. single word) correspondences.

The resulting bilingual dictionary contains 58,970 English-Dutch word pairs. More details can be found in table 2.5. A sample of the NL-Translex dictionary is shown in table 2.6.

	# Entries
English-Dutch word pairs	58,970
English words	43,914
Dutch words	43,292

Table 2.5: NL-Translex one-to-one correspondences.

[6] This work was carried out in the framework of the Stevin DPC project.

English word form	Dutch word form
affair	affair
affair	affaire
affair	avontuurtje
affair	boel
affair	zaak
affect	affect
affect	beïnvloeden
affect	raken
affectation	aanstellerij
affectation	affectatie
affectation	toewijzing
affected	aangeslagen
affected	aanstellerig
affected	geaffecteerd
affection	aanhankelijkheid
affection	affectie
affection	genegenheid

Table 2.6: Sample of the English-Dutch NL-Translex dictionary.

2.4.2 Statistical word alignment

Statistical word alignment tools determine translational correspondences at the level of the word, which can be used for bilingual dictionary extraction. Statistical word alignment is based on the assumption of co-occurrence: words that are translations of each other co-occur more often in aligned sentence pairs than that they occur randomly.

The most common approach to word alignment are generative models (e.g. the IBM translation models), which view the alignment process as the generation of a word in one language from another (Brown et al. 1993). The simplest IBM model – IBM Translation Model One – is a purely lexical model: it only takes into account the word frequencies of the source and target sentences. The higher numbered IBM Models build on IBM Model One and take into account word order (distortion) and model the probability that one source word aligns to more than one target word (fertility).

One of the shortcomings of the IBM models is that they only allow 1:n word mappings and are, hence, asymmetric. They can only model 1:n correspondences as they take the source word as their starting point and estimate conditional probabilities (i.e. the probability that a target word is a translation of a source word, given the source word). Multiword units (e.g. the Dutch word *regeringsleider* [En literal translation: *governmentleader*] corresponds with the English *Head of Government*) are problematic for the word-based models, as every word (*Head*, *of* and *Government*) is treated as a separate entry. To overcome this problem, the IBM models are often used in two directions: from source to target and from target to source.

A detailed description and comparison of the IBM models can be found in Och and Ney (2003). Their GIZA++ toolkit, which is part of Moses, an open source statistical machine translation system (Koehn et al. 2007), provides an efficient implementation of the IBM models.

Other approaches to word alignment include discriminative approaches and heuristic approaches. Discriminative feature-based models require annotated data for training. The main advantage of discriminative models is that it is easy to use additional features (e.g. part-of-speech codes, word association scores) that can help to improve the word alignment (Moore 2005, Liu and Lin 2005).

Heuristic approaches use statistical association measures to calculate word association scores between all words of the source and target sentence (Smadja, McKeown and Hatzivassiloglou 1996, Ker and Chang 1997). Heuristics (e.g. retain the alignments with the highest association scores) are then used to obtain the final word alignment.

In this thesis, we made use of the Perl implementation of IBM Model One that is part of the Microsoft Bilingual Sentence Aligner (Moore 2002) and of the GIZA++ toolkit (Och and Ney 2003).

2.4.3 Probabilistic bilingual lexicon

We derived bilingual dictionaries from the Model One output on different variants of the Europarl corpus (a 2.4, 5.6, and 9.3 million word subcorpus).

From the selected parts of the Europarl corpus, we removed all files that were manually aligned at sub-sentential level (see section 3.6). Also all files from the last quarter of 2000 were removed, as these data are used by the Machine Translation research community as test corpus.

All variants of the Europarl corpus were aligned at sentence level using the alignment tool that was released together with the Europarl corpus. The corpora were tokenized (see section 2.5) and converted to lowercase.

	2.4M	5.6M	9.3M
Aligned sentences	50,000	116,912	202,289
Total tokens	2,416,719	5,636,468	9,323,898
En word forms freq > 2	15,295	22,261	27,398
Nl word forms freq > 2	20,362	31,506	42,455
English-Dutch word pairs	16,728	21,486	28,342
English words	10,109	12,500	15,716
Dutch words	12,939	16,132	21,373

Table 2.7: Formal characterstics of the English-Dutch Europarl subcorpora used for dictionary creation (upper part) and of the resulting dictionary (lower part).

Table 2.7 gives an overview of the formal characteristics of the sentence-aligned parallel corpora that were used for deriving the bilingual dictionaries and the formal characteristics of the resulting Model One dictionaries.

Model One uses a parameter to determine the number of source words that are used during training. This parameter was set for each corpus so that all words that occurred at least twice in the corpus were taken into account.

As Model One is asymmetric, we ran the model in two directions: from English to Dutch and from Dutch to English. To get high-accuracy links, only the word pairs occurring in both the English-Dutch and Dutch-English word lists were retained, and the probabilities were averaged. To get rid of the noise produced by the translation model, only the entries with an averaged value of at least 0.1[7] were retained. This value was set experimentally.

To reduce the number of values, the averaged values were multiplied by 10 and only the integer part was retained. The obtained values are used to classify the different translations according to frequency.

A sample of the resulting dictionary is shown in table 2.8. As already mentioned, the dictionary does not abstract over word forms, e.g. *affordable - betaalbaar* and *affordable - betaalbare* are two separate entries in the dictionary. Model One can generate multiple translations for one word form, e.g. *affected* has three possible translations *getroffen, beïnvloed* and *getroffenen.*

[7] A probability of 0.1 approximately means that the words pair x-y co-occurred at least in 10% of the sentence pairs containing the target word y. This is an approximation as the model takes into account sentence length when estimating probabilities. The result of averaging the probability values of the English-Dutch and the Dutch-English translation models cannot be regarded as a probability anymore, as the number of occurrences of x and the number of occurrences of y in the corpus are different.

English word form	Dutch word form	Frequency class
aerospace	aerospace	2
affair	affaire	2
affairs	aangelegenheden	1
affairs	zaken	5
affect	invloed	1
affected	beïnvloed	1
affected	getroffen	4
affected	getroffenen	1
affection	genegenheid	1
affinity	affiniteit	2
affirmative	bevestigend	2
afflicted	geteisterd	1
afford	permitteren	3
afford	veroorloven	4
affordability	betaalbaarheid	2
affordable	betaalbaar	3
affordable	betaalbare	5
affords	biedt	1

Table 2.8: Sample of the English-Dutch Model One dictionary.

2.5 Shallow Linguistic Analysis

The sub-sentential alignment system that we have developed is chunk-driven and requires shallow linguistic processing tools for the source and target languages, i.e. part-of-speech taggers and chunkers. A lemmatizer is used to abstract over word forms before dictionary look-up.

Prior to adding additional linguistic information, the texts were segmented into sentences and tokenized:

- During sentence segmentation, sentence boundaries are marked using rules (e.g. *Put a sentence boundary after a sentence-final punctuation mark if followed by a word starting with a capital letter*) and constraints (e.g. *Do not put a sentence boundary after a full stop that is part of an abbreviation*).
- During tokenization, a sentence is split into sequences of words. All punctuation marks not belonging to the word form (i.e. punctuation marks that are not part of an abbreviation) are stripped off.

Sentence segmentation and tokenization for both English and Dutch were performed by a rule-based system using regular expressions and word lists. We used the tokenizer of the CNTS memory-based shallow parser developed by Sabine Buchholz and others.

The following example shows an English and Dutch tokenized sentence.

(1) En: She compared the appearance of Nefertiti 's mummy with the royal Egyptian fashion of that time and believes that the mummy can be identified as queen Nefertiti .

Nl: Ze vergeleek het uitzicht van Nefertiti's mummie met de koninklijke Egyptische mode uit die tijd en gelooft dat de mummie kan worden geïdentificeerd als koningin Nefertiti .

In the example a different treatment of the possessive marker *'s* in English and Dutch can be observed. According to the conventions of the English part-of-speech taggers, the possessive marker *'s* is split off during tokenization, and a separate PoS tag is assigned. According to the conventions of the Dutch part-of-speech tagger, the possessive marker is not stripped off during tokenization, and the possessiveness of the proper noun Nefertiti is encoded in the PoS tag.

Part-of-speech tagging and lemmatization for English was performed by the combined memory-based PoS tagger/lemmatizer, which is part of the MBSP tools (Daelemans and van den Bosch 2005). Part-of-speech tagging and lemmatization for Dutch was performed by TADPOLE (van den Bosch et al. 2007). Although the MBSP toolkits contain chunking for English and Dutch, we opted for the development of two rule-based chunkers, the reason being that the English and Dutch shallow parsers adopt a different chunk definition. For example, adjacent verbs are clustered in one verbal group in the English memory-based shallow parser, but regarded as separate chunks in the Dutch memory-based shallow parser.

2.5.1 Lemmatization

During lemmatization, for each orthographic token, the base form (lemma) is generated. The English and Dutch lemmatizers use similar definitions of base forms.

For verbs, this base form is the infinitive:

(2) En: believes - believe, identified - identify
Nl: vergeleek - vergelijken, gelooft - geloven, geïdentificeerd - identificeren

For most other words, this base form is the stem, i.e. the word form without inflectional affixes:

(3) En: scientists - scientist
Nl: koninklijke - koninklijk, archeologen - archeoloog

The lemmatizer makes use of the predicted PoS codes to disambiguate ambiguous word forms, e.g. Dutch *landen* can be an infinitive (base form *landen*) or plural form of a noun (base form *land*).

The English and Dutch lemmatizers, which are part of the MBSP tools, were trained on the English and Dutch parts of the Celex lexical database respectively (Baayen, Piepenbrock and van Rijn 1993).

2.5.2 Part-of-speech tagging

During part-of-speech tagging, a part-of-speech code is assigned to each orthographic token. The English and Dutch PoS taggers use different PoS tag sets.

The English memory-based PoS tagger was trained on data from the Wall Street Journal corpus of the Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993), and uses the Penn Treebank tag set. The Penn Treebank tag set contains 45 distinct tags (35 PoS tags and 10 other tags for punctuation marks and currency symbols).

Table 2.9 contains the English lemmata and PoS codes generated by the MBSP-tools for example sentence 1.

TADPOLE uses the CGN PoS tag set (Van Eynde, Zavrel and Daelemans 2000) adapted for written text (van den Bosch, Schuurman and Vandeghinste 2006). The CGN PoS tag set is characterized by a high level of granularity. Apart from the word class (e.g. noun, adjective, verb), the CGN PoS tag set codes a wide range of morpho-syntactic features (e.g. singular, plural, case information, tense) as attributes to the word class. In total, 316 distinct full tags are discerned.

Table 2.10 contains the Dutch lemmata and PoS codes for example sentence 1.

Token	Lemma	PoS
She	she	PRP
compared	compare	VBD
the	the	DT
appearance	appearance	NN
of	of	IN
Nefertiti	Nefertiti	NNP
’s	’s	POS
mummy	mummy	JJ
with	with	IN
the	the	DT
royal	royal	JJ
Egyptian	Egyptian	JJ
fashion	fashion	NN
of	of	IN
that	that	DT
time	time	NN
and	and	CC
believes	believe	VBZ
that	that	IN
the	the	DT
mummy	mummy	JJ
can	can	MD
be	be	VB
identified	identify	VBN
as	as	IN
queen	queen	NN
Nefertiti	Nefertiti	NNP
.	.	.

Table 2.9: English linguistic annotations.

Token	Lemma	PoS
Ze	ze	VNW(pers,pron,stan,red,3,ev,fem)
vergeleek	vergelijken	WW(pv,verl,ev)
het	het	LID(bep,stan,evon)
uitzicht	uitzicht	N(soort,ev,basis,onz,stan)
van	van	VZ(init)
Nefertiti's	Nefertiti	SPEC(deeleigen)
mummie	mummie	N(soort,ev,basis,zijd,stan)
met	met	VZ(init)
de	de	LID(bep,stan,rest)
koninklijke	koninklijk	ADJ(prenom,basis,met-e,stan)
Egyptische	egyptisch	ADJ(prenom,basis,met-e,stan)
mode	mode	N(soort,ev,basis,zijd,stan)
uit	uit	VZ(init)
die	die	VNW(aanw,det,stan,prenom,zonder,rest)
tijd	tijd	N(soort,ev,basis,zijd,stan)
en	en	VG(neven)
gelooft	geloven	WW(pv,tgw,met-t)
dat	dat	VG(onder)
de	de	LID(bep,stan,rest)
mummie	mummie	N(soort,ev,basis,zijd,stan)
kan	kunnen	WW(pv,tgw,ev)
worden	worden	WW(inf,vrij,zonder)
geïdentificeerd	identificeren	WW(vd,vrij,zonder)
als	als	VZ(init)
koningin	koningin	N(soort,ev,basis,zijd,stan)
Nefertiti	Nefertiti	SPEC(deeleigen)
.	.	LET()

Table 2.10: Dutch linguistic annotations.

To be able to compare the English and Dutch part-of-speech codes efficiently, a mapping table was defined. This mapping table is presented in appendix A.

2.5.3 Chunking

Chunking was introduced by Abney (1991) as a preliminary step towards full parsing, postponing more in-depth levels of analysis to a later phase. However, several language processing applications do not necessarily require full parses and can benefit from a superficial sentence analysis. The grouping of words into chunks or shallow parses has been succesfully used in various NLP applications such as speech synthesis (Buhmann et al. 2002), question answering (Buchholz and Daelemans 2001), and terminology extraction (Bourigault 1992). As shallow parsing only recovers a limited amount of syntactic information, shallow parsing is useful to analyze large volumes of text in an efficient and robust way (Daelemans and van den Bosch 2005, p. 85).

According to Abney, there is psychological evidence for the existence of chunks: chunks roughly correspond to prosodic units and are used by humans when processing sentences. During text chunking, syntactically related consecutive words are combined into non-overlapping, non-recursive chunks on the basis of a fairly superficial analysis.

Example 4 shows example sentence 1 divided in non-overlapping and non-recursive chunks.

(4) En: She | compared | the appearance | of Nefertiti 's mummy | with the royal Egyptian fashion | of that time | and | believes | that | the mummy | can be identified | as queen Nefertiti | .

Nl: Ze | vergeleek | het uitzicht | van Nefertiti's mummie | met de koninklijke Egyptische mode | uit die tijd | en | gelooft | dat | de mummie | kan worden geïdentificeerd | als koningin Nefertiti | .

Rule-based chunkers for Dutch and English were developed in the framework of this research project. These rule-based chunkers are based on a small set of constituency and distituency rules. Constituency rules define the part-of-speech tag sequences that can occur within a constituent (such as preposition + noun), while distituency rules define the part-of-speech tag sequences that cannot be adjacent within a constituent (such as noun + preposition). The use of distituency rules in the task of constituent boundary parsing was introduced by Magerman and Marcus (1990).

During development, the text files from the development corpus were used to fine-tune the constituency and distituency patterns. All English and Dutch text files from the development and test corpora (described in section 5.4.1) were manually chunked, and the rule-based chunkers were evaluated by running the CoNLL-evalscript developed by Tjong Kim Sang and Buchholz (2000) on the test files. Precision scores of 93% (English) and 94% (Dutch) and recall scores of 95% (English and Dutch) were obtained.

2.6 Summary

In this chapter, we dealt with the resources on which our sub-sentential system relies. Two types of bilingual data resources are used in our sub-sentential alignment system: sentence-aligned parallel corpora and bilingual dictionaries to retrieve lexical correspondences.

Apart from sentence-aligned parallel texts and a bilingual dictionary, our sub-sentential alignment system requires shallow linguistic processing tools for the source and target languages, i.e. part-of-speech taggers, lemmatizers and chunkers.

In the next chapter, we will have a closer look at the problem of translational correspondence and describe how a manual reference corpus was created.

CHAPTER 3

Manual Annotation of Translational Correspondence

3.1 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 where those complex translational correspondences are aligned manually is therefore a useful resource for studying shifts of translation, the linguistic changes that occur in the process of translating (Bakker et al. 2008).

In this chapter, we describe how we created a Gold Standard for the English-Dutch language pair[1]. 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 manually annotated corpora will be used throughout this thesis as test and development corpora.

[1] Part of the work described here is published in Macken (2010).

In section 3.2, an overview is given of related annotation projects. Section 3.3 sketches the problem of translational correspondence. Section 3.4 describes in detail how a Gold Standard for the English-Dutch language pair was created, and section 3.5 sets out how inter-annotator reliability was assessed. Section 3.6 elaborates on the corpus used.

3.2 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. 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, Andersson and Merkel 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 were manually provided. In the second case, all words in the text were 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 only be found 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). This 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 (2001b) 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 translation. In some annotation projects (Melamed 2001b, 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.

3.3 The problem of translational correspondence

The annotation task at hand is to link units of translational correspondence, i.e. units of the source text that correspond to an equivalent in the text of the translation. Barkhudarov defines this unit as:

> 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. (Barkhudarov 1993, p. 40)

These units can exist on any level of the language, and in most sentences translational correspondences are rather complex. Word-by-word correspondences can be found for some words only. The rest of the sentence is translated on the level of combinations of words. Translational mismatches of all kinds exist, even when the translations are semantically equivalent.

In the sections below, different types of translational correspondences are illustrated.

3.3.1 Word-by-word correspondence

The most straightfoward alignments are word-by-word correspondences. In the first example word-by-word correspondences can be indicated for all words, although the word order differs.

(5) En: The contract is worth several million euro .
Nl: Het contract is verscheidene miljoenen euro waard .

Slightly more complex correspondences occur when there are differences between source and target strings with respect to the use of grammatical function words. In the following example, the noun *productie* (En: *production*) is preceded by the definite article *de* (En: *the*) in Dutch. The translation also differs with respect to the tenses used, namely the present perfect *is...gestegen* [En: *has...risen*] versus the simple past *rose*.

(6) Nl: Gedurende dezelfde periode **is de productie gestegen** van 7 tot 11 miljoen ton .
En: Over the same period , **production rose** from 7 to 11 million tons .

A specific problem for the English-Dutch language pair is the different compounding strategy. Often, a single Dutch word (in the example *productinformatie*) corresponds to two or more English words (*product information*).

(7) En: The README.HTML file contains supplemental **product information** .
Nl: Het bestand README.HTML bevat aanvullende **productinformatie** .

3.3.2 Discontiguous correspondences

The corresponding units are not necessarily contiguous. Non-contiguous correspondences often occur when the source or the target sentence contains a flexible multiword expression, for example a separable or phrasal verb. Separable verbs (in the example *neemt...deel*) occur frequently in Dutch.

(8) En: Belgium **participates** in Eurovision Song contest after all .
Nl: België **neemt** uiteindelijk toch **deel** aan het Eurovisie Songfestival .

3.3.3 Phraseological units

A typical example of translation at the level of combinations of words are phraseological units. Phraseological units can be compounds, idioms, fixed expressions, multiword expressions, and specific terms. In most cases, the meaning of a phraseological unit cannot be derived from the (literal) meaning of the individual parts. In the example sentence *prominent* corresponds with the Dutch idiomatic expression *in het oog springend* [En literary translation: *in the eye jumping*].

(9) Nl: de meest **in het oog springende** delen van de installatie .
En: the most **prominent** parts of the setup .

3.3.4 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. Typical examples of this category are active-passive transformations. In the example below, the Dutch passive voice *worden benoemd* (En: *are appointed*) is translated by the active voice in English.

(10) Nl: De bestuurders **worden benoemd** door de algemene vergadering voor een termijn van vier jaar .
En: The General Meeting **appoints** the directors for a term of four years .

Greater structural discrepancies are found when lexical words are translated by a totally different syntactic construction. In the following example an English premodifier *last week's* is translated by a relative clause in Dutch *die vorige week hebben plaatsgevonden* (En literary translation: *that last week have taken place*).

(11) En: **last week's** attacks on innocent civilians
Nl: de aanvallen op onschuldige burgers **die vorige week hebben plaatsgevonden**

In some cases, the discrepancies are not only on the structural but also on the semantic level. In the following example *moord* (En: *murder*), is translated by *fatal incident* in English.

(12) Nl: Die **moord** riep een overreactie op van het Habsburgse rijk die leidde tot de Eerste Wereldoorlog .
En: That **fatal incident** elicited an overreaction , which led to the First World War .

3.3.5 Omissions and additions

In the translation process, the translator may have omitted or inserted some words. In the example below, *van het Habsburgse rijk* (En: *of the Habsburg empire*) is omitted in the English translation.

(13) Nl: Die moord riep een overreactie op **van het Habsburgse rijk** die leidde tot de Eerste Wereldoorlog .
En: That fatal incident elicited an overreaction , which led to the First World War .

3.4 Reference alignment

In order to create an a-priori reference alignment for a set of English-Dutch parallel texts, translational correspondences were indicated manually by a number of volunteers. To that end an annotation scheme was created and detailed annotation guidelines were written.

3.4.1 Annotation scheme

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.

3.4.2 Annotation guidelines

To improve consistency between the annotators, detailed annotation rules were written. The annotation manual is provided in appendix B. The annotation guidelines were updated regularly in the course of the annotation process, whenever new problematic cases were encountered.

The annotation guidelines are to a large extent based on the annotation guidelines of other word alignment projects (Melamed 2001b, Véronis 1998, Merkel 1999a). As a starting point, the Blinker project (Melamed 2001b) 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 in section 3.2, the Arcade project (Véronis 1998) and the Plug project (Merkel 1999a) were restricted to translation spotting: only for some given words was the translational correspondence in the target text 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 (1999a):

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

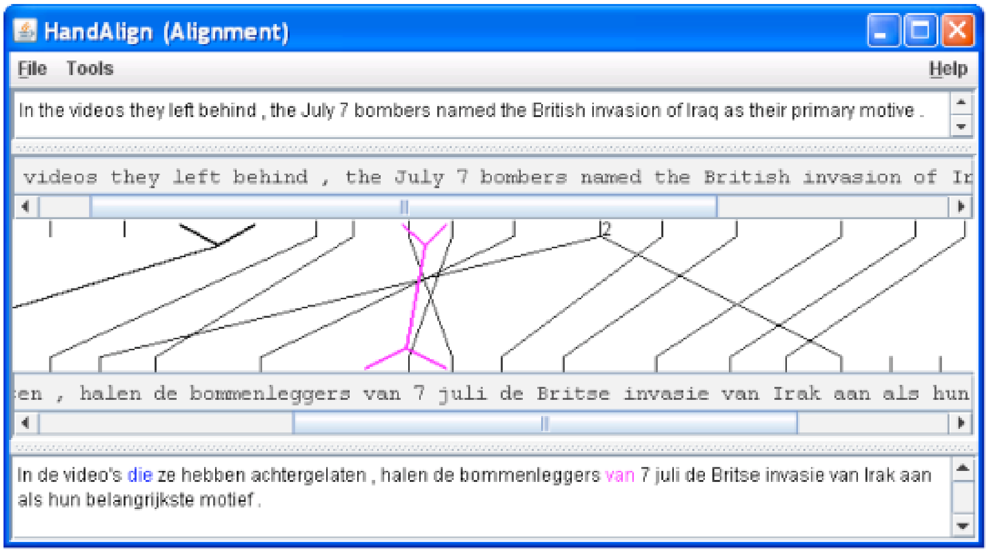
- Select *as few words as possible* in the source and in the target sentence, while preserving two-way equivalence.

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.

3.4.3 Annotation tool

To facilitate the annotation process, a graphical annotation tool, HandAlign[2], 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.


HandAlign (Alignment)
File
Tools
Help
videos they left behind , the July 7 bombers named the British invasion of Ir
2
en , halen de bommenleggers van 7 juli de Britse invasie van Irak aan als hun

In the videos they left behind , the July 7 bombers named the British invasion of Iraq as their primary motive .

In de video's die ze hebben achtergelaten , halen de bommenleggers van 7 juli de Britse invasie van Irak aan als hun belangrijkste motief .

Figure 3.1: Graphical annotation tool.

The annotator works in a graphical environment that consists of three panels (see figure 3.1):

- The top text area contains the source text.
- 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 (see section 2.5). 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.

[2] Available at http://www.cs.utah.edu/~hal/HandAlign/

For further processing, the output of HandAlign was converted into a table. In the table representation (see table 3.1), the first column contains the source text segment, the second column the target text segment, and the third column the type of link (0 = null link, 1 = fuzzy link, 2 = 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
In	In	2
the	de	2
videos	video's	2
null	die	0
they	ze	2
left+behind	hebben+achtergelaten	2
,	,	2
the	de	2
July+7	van+7+juli	1
July	juli	2
7	7	2
bombers	bommenleggers	2
named	halen+...+aan	2
the	de	2
British	Britse	2
invasion	invasie	2
of	van	2
Iraq	Irak	2
as	als	2
their	hun	2
primary	belangrijkste	2
motive	motief	2
.	.	2

Table 3.1: Table representation of the manual alignments (0 = null link, 1 = fuzzy link, 2 = regular link).

3.5 Inter-annotator reliability

In order to assess the feasibility of the alignment task given the annotation guidelines, an inter-annotator experiment was set up. 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.

3.5.1 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 3.2. 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 exhibit 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.

3.5.2 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:

$$WAA = \frac{Weight_{Agree}}{Weight_{Total}} \tag{3.1}$$

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 (2001c)[3].

3.5.3 Inter-annotator results

Detailed results for all test files are presented in table 3.2. Apart from the Kappa score and WAA score, the statistics include 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		Agreement	
Text	Ann1	Ann2	Ann1	Ann2	Ann1	Ann2	Kappa	WAA
2000-03-24	88.4	87.2	6.9	7.6	4.8	5.2	0.73	0.86
2001-03-26	87.9	89.8	6.7	4.3	5.4	5.8	0.80	0.86
2001-05-16	93.1	90.8	3.2	4.8	3.7	4.5	0.90	0.94
2002-10-08	85.1	88.0	8.9	5.9	6.0	6.2	0.71	0.86
2003-08-27	94.7	93.2	2.3	2.7	3.0	4.1	0.83	0.92
2005-01-26b	91.7	93.3	5.0	3.6	3.3	3.1	0.79	0.89
2005-05-04	90.6	89.2	7.5	6.8	1.9	4.1	0.73	0.84
2005-09-30	94.0	93.6	4.4	3.7	1.7	2.7	0.73	0.94

Table 3.2: Overview of the inter-annotator alignment scores: percentage of regular, fuzzy and null links used by each annotator, Kappa score and WAA score.

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 whole corpus.

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

3.6 Corpus

As was explained in section 3.1, our Gold Standard data set contains texts from different text types in order to cover a wide range of syntactic and stylistic phenomena that emerge from a different writing and translation style. The reference corpus consists of six different text types: journalistic texts, proceedings of plenary debates, press releases, newsletters, medical European Public Assessment Reports and user manuals. We assume that for each of the six text types another translation style was adopted, with the journalistic texts being the most free translations and the user manuals being the most literal translations.

Table 3.3 summarizes the formal characteristics of the corpus. In total, the Gold Standard contains more than 58,000 words.

Text type	Total Words	Sentence length (source)	Sentence length (target)	Ratio S/T sentences
Journalistic texts	7,706	22.0	20.0	0.88
Proceedings EP	9,463	24.8	21.1	0.78
Press releases	13,871	24.6	23.7	0.97
Newsletters	10,480	15.0	15.4	0.99
EPARs	7,536	17.2	17.7	1.01
User manuals	9,558	13.1	13.1	1.00

Table 3.3: Corpus characteristics of the Gold Standard.

3.6.1 Journalistic texts

The journalistic texts were extracted from the Dutch Parallel Corpus (see section 2.2.3). The 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 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.

3.6.2 Proceedings of plenary debates

The Europarl corpus (Koehn 2005) contains the proceedings of the European Parliament plenary debates in each of the eleven official languages of the European Union. The target audience of the proceedings are all European citizens. The texts are a written reproduction of prepared speeches.

As explained in section 2.2.1, an English-Dutch subcorpus of 6 million words of direct translations was extracted from Europarl. From this subcorpus, texts were selected from the debates in the days following the 9/11-attacks of 2001.

The English source sentences are relatively long, with an average sentence length of 24.8 words. The ratio source/target sentences is 0.78, which means that a considerable number of 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 (21.1 words vs. 24.8 words).

3.6.3 Press releases

The press releases were provided by Barco, a Belgian high-tech company in the visualization industry. The texts were selected from the company's archive of press releases on the basis of content: all texts describe the use of the company's technology on worldwide events, or announce new partnerships. Due to the nature of the company, the press releases contained a high percentage of technical terms. The texts were originally written in English and translated into Dutch.

The average sentence length is 24.6 words for the English source texts and 23.7 words for the Dutch target texts. The ratio source/target sentences is 0.97, which means that only a few source sentences are translated by two or more target sentences.

3.6.4 Newsletters

The newsletters were extracted from the Dutch Parallel Corpus. They 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.

3.6.5 European Public Assessment Reports

The European Public Assessment Reports (EPARs) are part of the Dutch Parallel Corpus. 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.

3.6.6 User manuals

The user manuals come from three different companies: SAP[4], IBM and Philips. These technical documents are typically written for an end user. The texts are characterized by a high percentage of technical terms. The sentences are short, with an average of 13 words per sentence. Mathematically, there is a perfect one-to-one correspondence between source and target sentences.

3.6.7 Overview of links

In table 3.4 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 links and null links.

The journalistic texts and the proceedings of the debates contain the highest number of fuzzy links (11 and 10% respectively) and the highest number of null

[4] The SAP files are part of the training kits of the EColoRe project, available at http://ecolore.leeds.ac.uk.

links (9 and 8% respectively). The user manuals contain the lowest percentage of fuzzy links (5%), and very few null links (2.7%). The newsletters, press releases and EPARs are somewhere in between.

A detailed analysis of the manually indicated translational correspondences of the proceedings of the debates, the press releases and the user manuals can be found in Macken (2007).

Text type	Regular	Fuzzy	Null
Journalistic texts	79.6	11.1	9.3
Proceedings EP	81.6	10.3	8.1
Press releases	89.3	6.0	4.7
Newsletters	88.6	6.9	4.5
EPARs	89.6	6.5	3.9
User manuals	92.0	5.3	2.7

Table 3.4: Percentage of regular, fuzzy and null links.

3.7 Summary

In this chapter, 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 indicated: 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.

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.

The results of the inter-annotator experiment seemed sufficiently high to apply the annotation procedure on the whole corpus..

In order to cover a wide range of syntactic and stylistic phenomena that emerge from a different writing and translation style, 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 manual alignments of the journalistic texts, the newsletters and the EPARs are publicly available as part of the Dutch Parallel Corpus.

In the following chapter, we will explain how this Gold Standard data set is used to assess the performance of automatic alignment systems.

CHAPTER 4

Evaluation metrics

4.1 Introduction

This chapter deals with evaluation. In order to be able to compare different alignment systems, some objective means of evaluating their performance is required. The evaluation of word or chunk alignment systems is not a trivial task. In this chapter we discuss and compare the different evaluation metrics that will be used throughout this thesis.

In the previous chapter we described how a Gold Standard was created. As translational correspondences are often difficult to determine at word level, different units could be linked. Three types of links were introduced: *regular* links to connect straightforward correspondences; *fuzzy* links for translation-specific shifts of various kinds; and *null* links for source text units that have not been translated or target text units that have been added. A multi-level annotation was used in the case of divergent translations: fuzzy links were used to connect paraphrased sections, regular links were used to connect corresponding words within the paraphrased sections.

If we want to compare the output of an automatic alignment system with the manually created Gold Standard, we are faced with two – somewhat related – problems: the first problem is that the alignment system can align different units; the second problem is that the proposed alignments can be only partially correct.

To solve the first problem, all word-based evaluation metrics use the technique that was introduced in the inter-annotator experiment described in chapter 3, and convert 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. Table 4.1 shows the resulting word-to-word alignments for the manual reference example that was presented in table 3.1 in chapter 3.

We repeat the sentence pair in example 14. In the case of a multi-level annotation, the regular word-to-word links overwrite the fuzzy word-to-word links. In the example, *July 7* is linked as a divergent phrase translation to *van 7 juli* [En: *of 7 July*]. In the conversion step, each word of the source phrase is interlinked with each word of the target phrase. The regular word-to-word links (*July* $\sim$ *juli* and *7* $\sim$ *7*) are labelled as regular link, the other resulting word-to-word alignments are labelled as fuzzy links.

(14) En: In the videos they left behind , the July 7 bombers named the British invasion of Iraq as their primary motive .

Nl: In de video's die ze hebben achtergelaten , halen de bommenleggers van 7 juli de Britse invasie van Irak aan als hun belangrijkste motief .

The different evaluation metrics mainly differ in the way that divergent translational correspondences are treated and how they deal with partially correct alignments. All metrics make use of the notion of precision and recall, which are widely used in the context of information retrieval:

- **Precision** measures how many links generated by the system are correct.
- **Recall** measures how many links were found by the system; it is thus a measure of the coverage of the system.

In the sections below, we will discuss four evaluation metrics in detail: F-measure, Alignment Error Rate (AER), a weighted version of F-measure, and F-measure calculated at chunk level.

Source word	Target word	Type of link
In	In	R
the	de	R
videos	video’s	R
∅	die	N
they	ze	R
left	hebben	R
left	achtergelaten	R
behind	hebben	R
behind	achtergelaten	R
,	,	R
the	de	R
July	van	F
July	7	F
July	juli	R
7	van	F
7	7	R
7	juli	F
bombers	bommenleggers	R
named	halen	R
named	aan	R
the	de	R
British	Britse	R
invasion	invasie	R
of	van	R
Iraq	Irak	R
as	als	R
their	hun	R
primary	belangrijkste	R
motive	motief	R
.	.	R

Table 4.1: Manual reference represented as word-to-word alignments (’R’ stands for regular link; ’F’ for fuzzy link and ’N’ for null link).

4.2 F-measure

A straightforward measure to use is F-measure, which combines precision and recall calculated on all word-to-word links.

To calculate precision, recall and F-measure on all word-to-word links, the following equations were used (R refers to the set of word-to-word alignments of the manually created reference alignment and A refers to the set of alignments generated by the system):

$$Precision = \frac{|A \cap R|}{|A|} \tag{4.1}$$

$$Recall = \frac{|A \cap R|}{|R|} \tag{4.2}$$

$$F{-}measure = \frac{2 * Precison * Recall}{Precision + Recall} \tag{4.3}$$

We will calculate F-measure for the example presented in table 4.2 below. The left-hand side of this table contains all the reference word-to-word links; the right-hand side contains all the word-to-word links generated by an automatic alignment system.

The manual reference alignment contains 30 word-to-word alignments. The alignment system proposes 23 word-to-word alignments of which two are wrong (*behind* ∼ *van* and *named* ∼ *bommenleggers*). The system missed 9 word-to-word links: 1 null link (*die*), 4 word-to-word links derived from two regular phrasal alignments (*behind* ∼ *hebben*, *behind* ∼ *achtergelaten*, *named* ∼ *halen* and *named* ∼ *aan*) and 4 word-to-word links derived from one fuzzy phrasal alignment(*July* ∼ *van*, *July* ∼ *7*, *7* ∼ *van* and *7* ∼ *juli*).

For this example, the obtained precision score is 0.91 (21 / 23), the recall score 0.70 (21 / 30) and the F-measure 0.79 ((2 * 0.91 * 0.7) / (0.91 + 0.70)).

4.3 Alignment Error Rate

Alignment Error Rate (AER) was introduced by Och and Ney (2003). In their manual reference alignment, they distinguished *sure* alignments (S) and *possible* alignments (P) (see section 3.2) and introduced the following redefined precision

Reference			System	
Source word	Target word	Type of link	Source word	Target word
In	In	R	In	In
the	de	R	the	de
videos	video's	R	videos	video's
∅	**die**	N		
they	ze	R	they	ze
left	hebben	R	left	hebben
left	achtergelaten	R	left	achtergelaten
behind	**hebben**	R		
behind	**achtergelaten**	R		
			behind	**van**
,	,	R	,	,
the	de	R	the	de
July	**van**	F		
July	**7**	F		
July	juli	R	July	juli
7	**van**	F		
7	7	R	7	7
7	**juli**	F		
bombers	bommenleggers	R	bombers	bommenleggers
named	**halen**	R		
named	**aan**	R		
			named	**bommenleggers**
the	de	R	the	de
British	Britse	R	British	Britse
invasion	invasie	R	invasion	invasie
of	van	R	of	van
Iraq	Irak	R	Iraq	Irak
as	als	R	as	als
their	hun	R	their	hun
primary	belangrijkste	R	primary	belangrijkste
motive	motief	R	motive	motief
.	.	R	.	.

Table 4.2: Reference word-to-word links and word-to-word links generated by an automatic alignment system. Missed and erroneous alignments are indicated in bold.

and recall measures (where A refers to the set of alignments generated by the system):

$$Precision_{AER} = \frac{|A \cap P|}{|A|} \tag{4.4}$$

$$Recall_{AER} = \frac{|A \cap S|}{|S|} \tag{4.5}$$

and the Alignment Error Rate (AER):

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|} \tag{4.6}$$

The distinction between *sure* and *possible* alignments approximately corresponds to the distinction between regular and fuzzy links in our annotation scheme. Therefore we consider all regular links of the manual reference as *sure* alignments and all fuzzy and null links as *possible* alignments to compare the output of an alignment system with the manual reference alignments.

For the example presented in table 4.2, the obtained AER precision score is 0.91 (21 / 23), the AER recall score is 0.84 (21 / 25) and the AER is 0.125 (1 - (21 + 21) / (23 + 25)).

4.4 Weighted F-measure

Weighted F-measure was introduced by Melamed (2001c) to adjust for all word-to-word links resulting from phrasal alignments. In the F-measure presented in section 4.2, all word-to-word links are equally important. However, as Melamed (2001c) pointed out, an evaluation metric that treats all links as equally important would place undue importance on words that were linked more than once. Therefore, in the weighted version of F-measure, a weight is assigned to each word-to-word link, and precision and recall are calculated on the normalized weights of all word-to-word links.

We use the weighting method developed by Davis, Xie and Small (2007), which is a refinement of the weighting principles introduced by Melamed (2001c)[1]. In this weighting scheme, every word contributes 0.5 to the total weight. In the

[1] Melamed goes through several steps of weighting and re-weighting links. The method of Davis et al. (2007) is much simpler to compute and serves the same purpose.

Reference			System		
Source word	Target word	Weight	Source word	Target word	Weight
In	In	1.00	In	In	1.00
the	de	1.00	the	de	1.00
videos	video’s	1.00	videos	video’s	1.00
∅	**die**	0.50			
they	ze	1.00	they	ze	1.00
left	hebben	0.50	left	hebben	0.75
left	achtergelaten	0.50	left	achtergelaten	0.75
behind	**hebben**	0.50			
behind	**achtergelaten**	0.50			
			behind	**van**	1.00
,	,	1.00	,	,	1.00
the	de	1.00	the	de	1.00
July	**van**	0.416			
July	**7**	0.416			
July	juli	0.416	July	juli	1.00
7	**van**	0.416			
7	7	0.416	7	7	1.00
7	**juli**	0.416			
bombers	bommenleggers	1.00	bombers	bommenleggers	0.75
named	**halen**	0.75			
named	**aan**	0.75			
			named	**bommenleggers**	0.75
the	de	1.00	the	de	1.00
British	Britse	1.00	British	Britse	1.00
invasion	invasie	1.00	invasion	invasie	1.00
of	van	1.00	of	van	1.00
Iraq	Irak	1.00	Iraq	Irak	1.00
as	als	1.00	as	als	1.00
their	hun	1.00	their	hun	1.00
primary	belangrijkste	1.00	primary	belangrijkste	1.00
motive	motief	1.00	motive	motief	1.00
.	.	1.00	.	.	1.00
Total Weight		23.5			22

Table 4.3: Reference word-to-word links and word-to-word links generated by an automatic alignment system and their normalized weights. Missed and erroneous alignments are indicated in bold.

case of interlinked word-to-word links from the phrasal alignments, each link is assigned the total weight of the phrasal alignment divided by the number of word-to-word links.

In table 4.3, the normalized weights are presented for the manual reference alignments and the system alignments. In the example *named* ∼ *halen aan*, the total weight of the phrasal alignment is 1.5 (= 3 x 0.5), and each word-to-word link gets a weight of 0.75 (= 1.5 / 2). In the case of a one-to-one word-to-word link (e.g. *invasion* ∼ *invasie*), the resulting weight of the word-to-word link is 1 (2 x 0.5 / 1). Null links get a weight of 0.5.

Precision, recall and F-measure are then calculated on the normalized weights, by using the following equations:

$$Precision_{Weighted} = \frac{Weight_A(A \cap R)}{Weight_A(A)} \tag{4.7}$$

$$Recall_{Weighted} = \frac{Weight_R(A \cap R)}{Weight_R(R)} \tag{4.8}$$

$$F{-}measure_{Weighted} = \frac{2 * Precison_{Weighted} * Recall_{Weighted}}{Precision_{Weighted} + Recall_{Weighted}} \tag{4.9}$$

For the example presented in table 4.3, the obtained weighted precision score is 0.92 (20.25 / 22), the weighted recall score is 0.82 (19.336 / 23.5) and the weighted F-measure is 0.87 ((2 * 0.92 * 0.82) / (0.92 + 0.82)).

4.5 F-measure calculated at chunk level

All evaluation metrics presented above operate at the word level. In the next chapter, we will explain that our sub-sentential alignment system is chunk-based and links chunks rather than words. Therefore, we also calculate precision and recall on the level of aligned chunks.

On the basis of the established word alignments and the chunk boundaries, we derive all aligned chunks. Table 4.4 shows the aligned chunks for our example. The upper part of the table contains all manually aligned chunks. The lower part of the table contains the chunks that were aligned by an automatic alignment system.

For the chunk-based evaluation, precision, recall and F-measure are then calculated on the aligned chunks. For the example presented in table 4.4, the obtained precision score at chunk level is 0.78 (7/9), the recall score at chunk level is 0.64 (7/11) and the F-measure at chunk level is 0.70 ((2 * 0.78 * 0.64) / (0.78 + 0.64)).

As can be seen in table 4.4, F-measure calculated at chunk level penalizes erroneous word-to-word links more severely than the word-based metrics as an erroneous link (*behind* ∼ *van*) might have an impact on more than one aligned chunk (*left* ∼ *hebben achtergelaten* and *behind...the July 7 bombers named* ∼ *de bommenleggers van 7 juli*).

Source chunk(s)	Target chunk(s)
REFERENCE	
In the videos	In de video's
∅	**die**
they	ze
left behind	**hebben achtergelaten**
,	,
the July 7 bombers	**de bommenleggers van 7 juli**
named	**halen ... aan**
the British invasion	de Britse invasie
of Iraq	van Irak
as their primary motive	als hun belangrijkste motief
.	.
SYSTEM	
In the videos	In de video's
they	ze
left	**hebben achtergelaten**
behind ... the July 7 bombers named	**de bommenleggers van 7 juli**
,	,
the British invasion	de Britse invasie
of Iraq	van Irak
as their primary motive	als hun belangrijkste motief
.	.

Table 4.4: Aligned chunks in the manual reference and generated by an automatic alignment system. Missed and erroneous chunks alignments are indicated in bold.

4.6 Summary

We presented three widely-accepted word-based evaluation metrics for word alignment. Especially for complex and divergent translational correspondences, comparing and scoring different alignment systems is not a trivial task, as these alignments cannot be simply classified as *right* or *wrong.* Two different solutions for scoring complex translational correspondences were presented: a weighted version of F-measure and AER. Weighted F-measure focuses on the correct treatment of phrasal alignments and the related problem of partial correctness, while AER makes use of the distinction between regular and fuzzy links.

In the next chapters, we will use weighted F-measure as our standard evaluation metric. In chapter 6, we will present the results both in terms of weighted F-measure and AER in order to be able to compare our results with other alignment projects reported in literature. We will demonstrate that, although AER cannot be compared directly with weighted F-measure, weighted F-measure and AER do correlate with each other.

Weighted F-measure and AER are global metrics in the sense that they combine precision and recall. However, since this thesis focuses on precision, we want to be able to assess the precision of different alignment systems. For that reason, we do not only present weighted F-measure and AER but also the underlying precision and recall scores.

As the system we developed is chunk-based, we also calculate – apart from the word-based metrics – F-measure at chunk level. When interpreting the results, however, we keep in mind that F-measure at chunk level is very severe and does not account for partially correct chunk alignments.

CHAPTER 5

Anchor Chunk Alignment

5.1 Introduction

Thus far, we have described all the preliminary work that was necessary for the creation of our sub-sentential alignment system. In chapter 2, we elaborated on all the resources on which our system relies: sentence-aligned parallel corpora, a bilingual dictionary or statistical word alignment tools, and shallow linguistic processing tools. In chapter 3, we introduced the manually annotated corpora that will be used throughout this thesis as test and training corpora, and in chapter 4 we discussed the different evaluation metrics that will be used.

We now arrive at the core part of this thesis, viz. the description of our sub-sentential alignment system. In this chapter, we present the general architecture of our sub-sentential alignment system that links linguistically motivated phrases in parallel texts and we present in detail the *Anchor Chunk Alignment* step[1].

The basic idea behind our approach is that – at least for European languages

[1] This chapter is a more elaborate version of work published in Macken and Daelemans (2009).

– translations conveying the same meaning use to a certain degree the same building blocks from which this meaning is composed: i.e. we assume that to a large extent noun phrases and prepositional phrases, verb phrases and adverbial phrases in one language directly map to similar constituents in the other language. The extent to which our basic assumption holds depends on the translation strategy that was used. Text types that are typically translated in a more literal way (e.g. technical texts) will contain more direct correspondences than text types for which a freer translation strategy was adopted (e.g. journalistic texts).

This underlying assumption is reflected in the global architecture of our sub-sentential alignment system. We conceive our sub-sentential aligner as a cascaded model consisting of two phases. In the first phase *anchor chunks*, i.e. chunks that can be linked with a very high precision on the basis of lexical correspondences and syntactic similarity are retrieved. In the second phase, we focus on the more complex translational correspondences on the basis of observed translation shift patterns. The anchor chunks of the first phase are used to limit the search space in this second phase. This chapter describes the general architecture and the first phase, namely the alignment of anchor chunks. The next chapter deals with the second phase of the process.

The sub-sentential alignment module has been implemented in Java. The most important classes and methods are described in appendix C.

5.2 Architecture

The global architecture of our system is visualized in figure 5.1. The sub-sentential alignment system takes as its input sentence-aligned texts, together with additional linguistic annotations (part-of-speech codes and lemmata) for the source and target text.

Although the global architecture of our sub-sentential alignment system is language-independent, some language-specific resources are used. In the first phase, two *external* language-specific linguistic resources are needed: first, a bilingual lexicon to generate the lexical correspondences; second, tools to generate additional linguistic information (part-of-speech tagger, lemmatizer and chunker).

We experimented with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries, automatically extracted from bilingual corpora of various sizes.

As explained in chapter 2 (section 2.5), part-of-speech tagging and lemmati-

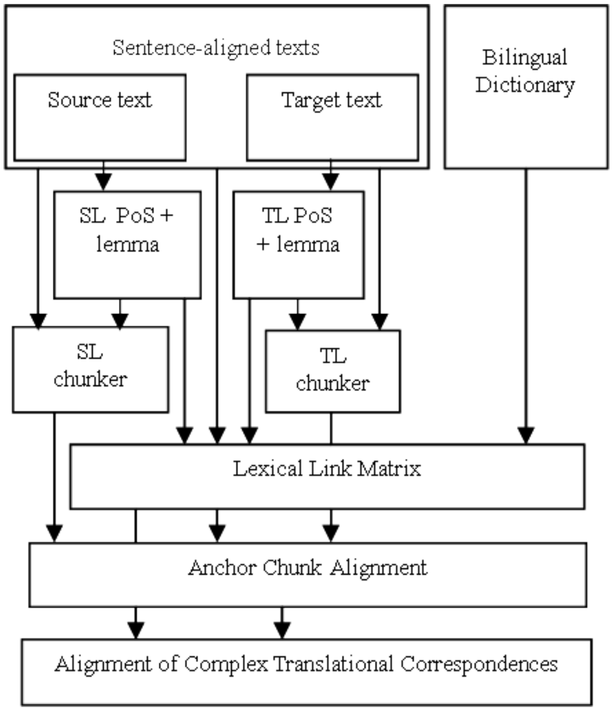


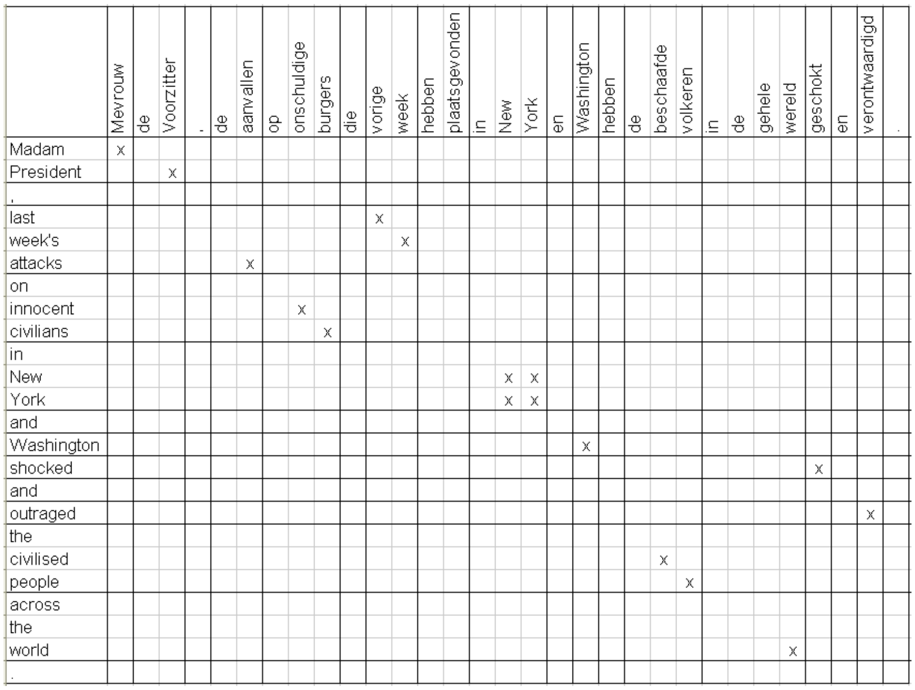
Figure 5.1: Outline of the full sub-sentential alignment system.

zation for English was performed by the combined memory-based PoS tagger/lemmatizer, which is part of the MBSP tools (Daelemans and van den Bosch 2005). Part-of-speech tagging and lemmatization for Dutch was performed by TADPOLE (van den Bosch et al. 2007). The rule-based chunkers developed in the framework of this PhD project were used to generate the chunk boundaries for Dutch and English (see section 2.5.3).

The different steps of the sub-sentential alignment process are simulated in figures 5.2 and 5.3 for the following example sentence pair:

(15) En: Madam President, last week's attacks on innocent civilians in New York and Washington shocked and outraged the civilised people across the world.

Nl: Mevrouw de Voorzitter, de aanvallen op onschuldige burgers die vorige week hebben plaatsgevonden in New York en Washington hebben de beschaafde volkeren in de gehele wereld geschokt en verontwaardigd.



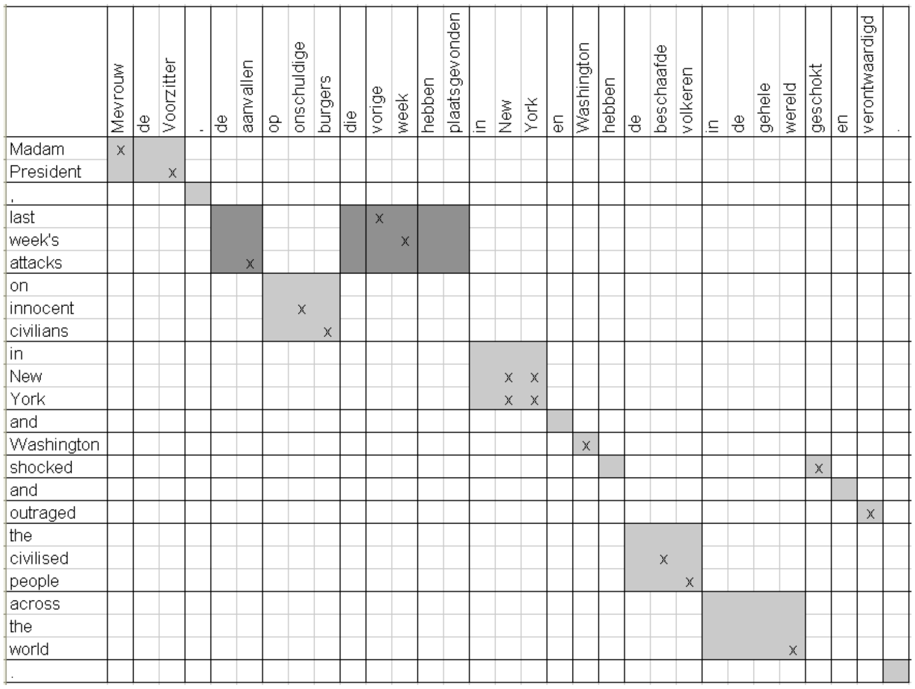
	Mevrouw	de	Voorzitter	,	de	aanvallen	op	onschuldige	burgers	die	vorige	week	hebben	plaatsgevonden	in	New	York	en	Washington	hebben	de	beschaafde	volkeren	in	de	gehele	wereld	geschokt	en	verontwaardigd	.
Madam	x																													
President			x																											
,																														
last											x																			
week's												x																		
attacks						x																								
on																														
innocent								x																						
civilians									x																					
in																														
New																x	x													
York																x	x													
and																														
Washington																			x											
shocked																												x		
and																														
outraged																														x
the																														
civilised																						x								
people																							x							
across																														
the																														
world																											x			
.																														

Figure 5.2: Simulation of the different alignment steps. Chunk boundaries are indicated by horizontal and vertical lines. Lexical correspondences are indicated by x's.

In the first step of the process, the source and target sentences are divided into chunks on the basis of part-of-speech information, and lexical correspondences are retrieved from a bilingual dictionary. The chunk boundaries are visualized in figure 5.2 by means of horizontal and vertical lines. The lexical correspondences are marked by x's.

During anchor chunk alignment, the sub-sentential aligner links chunks on the basis of lexical correspondences and chunk similarity. In figure 5.3, the anchor chunks are marked in light grey. In the example sentence, corresponding noun phrases, corresponding prepositional phrases and corresponding verb phrases are indicated.

In the second phase of the process, the sub-sentential aligner uses the anchor chunks of the first phase to retrieve the more complex translational correspondences. In figure 5.3, such a more complex translational correspondence is marked in dark grey. In the example sentence, the following translation shift can be observed: the English premodifier (*last week's*) is translated by a relative clause in Dutch (*die vorige week hebben plaatsgevonden*).



	Mevrouw	de	Voorzitter	,	de	aanvallen	op	onschuldige	burgers	die	vorige	week	hebben	plaatsgevonden	in	New	York	en	Washington	hebben	de	beschaafde	volkeren	in	de	gehele	wereld	geschokt	en	verontwaardigd	.
Madam	x																														
President			x																												
,																															
last											x																				
week's												x																			
attacks						x																									
on																															
innocent								x																							
civilians									x																						
in																															
New																x	x														
York																x	x														
and																															
Washington																			x												
shocked																												x			
and																															
outraged																														x	
the																															
civilised																						x									
people																							x								
across																															
the																															
world																											x				
.																															

Figure 5.3: Simulation of the different alignment steps. Anchor chunks are marked in light grey. Complex translational correspondence are marked in dark grey.

5.3 Algorithm

As explained in section 5.2, we conceive our sub-sentential alignment system as a cascaded model consisting of two phases. The objective of the first phase is to link *anchor chunks*, i.e. chunks that can be linked with a very high precision. Those anchor chunks are linked on the basis of lexical clues and chunk similarity.

In order to link chunks on the basis of lexical clues and chunk similarity, the following steps are taken for each sentence pair:

1. Creation of a lexical link matrix.

2. Linking chunks on the basis on lexical correspondences and chunk similarity.

3. Linking adjacent function word chunks and final punctuation.

The different steps are described in more detail below.

5.3.1 Creation of the lexical link matrix

Prior to creating the lexical link matrix, all possible translations for each word in the source and target sentence are retrieved from a bilingual dictionary. As explained in section 2.4.3, the probabilistic bilingual dictionary contains English-Dutch word pairs with numeric values that denote the frequency class of the word pair[2].

For each source and target word, all translations for the word form and the lemma are retrieved from the bilingual dictionary. If only the lemma of the source or target word is found in the bilingual dictionary, the resulting frequency weight is cut in half.

In the process of building the lexical link matrix, function words are neglected. Given the frequency of function words in a sentence, linking function words based on word alignment information alone often results in erroneous alignments. For that reason no lexical links are created for the following word classes: determiners, prepositions, coordinating conjunctions, possessive pronouns and punctuation symbols.

For all content words, if a source word occurs in the set of possible translations of a target word, or if a target word occurs in the set of possible translations of the source words, a lexical link is created. Identical strings in the source and the target sentence are also linked.

The resulting lexical link matrix for our example is shown in figure 5.4.

5.3.2 Linking anchor chunks

The problem of linking chunks based on lexical correspondences and similar chunks can be decomposed in two sub-problems:

[2] As in the NL-Translex dictionary no frequency information is available, all word pairs get the same value.

	Mevrouw	de	Voorzitter	,	de	aanvallen	op	onschuldige	burgers	die	vorige	week	hebben	plaatsgevonden	in	New	York	en	Washington	hebben	de	beschaafde	volkeren	in	de	gehele	wereld	geschokt	en	verontwaardigd	.
Madam	4		1																												
President			5																												
,																															
last											3																				
week's												1																			
attacks						4																									
on																															
innocent								5																							
civilians									2																						
in																															
New																9															
York																2	9														
and																															
Washington																			9												
shocked																												5			
and																															
outraged																															
the																															
civilised																						3									
people																															
across																															
the																															
world																											7				
.																															

Figure 5.4: Lexical link matrix containing frequency weights for example 15.

1. Selecting candidate anchor chunks.
2. Testing chunk similarity of the candidate anchor chunks.

Selecting candidate anchor chunks

The candidate anchor chunks are selected on the basis of the information available in the lexical link matrix. For each source chunk a **contiguous** candidate target chunk is constructed. The contiguous candidate target chunk is built by concatenating all target chunks from a *begin index* until an *end index*. The begin index points to the first target chunk with a lexical link to the source chunk under consideration. The end index points to the last target chunk with a lexical link to the source chunk under consideration. Possible intermediate chunks can contain additional lexical links, but this is not necessarily the case. If a source word contains more than one lexical link, the lexical link with the highest frequency weight is used.

In this way, the following contiguous 1:1 and 1:n candidate target chunks are built for our example:

<table>
<tr><td>Madam President</td><td>Mevrouw | de Voorzitter</td></tr>
<tr><td>last week's attacks</td><td>de aanvallen | op onschuldige burgers | die | vorige week</td></tr>
<tr><td>on innocent civilians</td><td>op onschuldige burgers</td></tr>
<tr><td>in New York</td><td>in New York</td></tr>
<tr><td>Washington</td><td>Washington</td></tr>
<tr><td>shocked</td><td>geschokt</td></tr>
<tr><td>the civilised people</td><td>de beschaafde volkeren</td></tr>
<tr><td>across the world</td><td>in de gehele wereld</td></tr>
</table>

For some source chunks, it is also useful to build a **non-contiguous** candidate target chunk. The non-contiguous candidate target chunks are built by concatenating all target chunks with a lexical link to the source chunk under consideration. In our example, only one non-contiguous target chunk is constructed:

<table>
<tr><td>last week's attacks</td><td>de aanvallen . . . vorige week</td></tr>
</table>

The process of selecting candidate chunks as described above, is performed twice: a first time starting from the source sentence; a second time starting from the target sentence. It is necessary to start the selection process from the target sentence to build n:1 candidate pairs like the following:

<table>
<tr><td>the whole | of the European Union</td><td>de gehele Europese Unie</td></tr>
</table>

In the second loop, only those chunks for which no similarity test was performed are taken into consideration.

Testing chunk similarity

For each selected candidate pair, a *similarity test* is performed. Chunks are considered to be similar if at least a certain percentage of words of source and target chunk(s) are either linked by means of a lexical link or can be linked on the basis of corresponding part-of-speech codes.

All word classes can be linked on the basis of PoS codes. In the candidate anchor chunk *the civilised people - de beschaafde volkeren*, one lexical clue (*civilised - beschaafde*) is sufficient to pass the similarity test because *the* and *de*, and *people* and *volkeren* are linked on the basis of corresponding PoS codes.

The percentage of words that have to be linked was empirically set at 80% for contiguous chunks and 100% for non-contiguous chunks. The percentage of linked words is calculated according to the following equation (where linked$_{src}$ refers to the number of linked words of the source chunk, linked$_{tgt}$ refers to the number of linked words of the target chunk, total$_{src}$ to the total number of words of the source chunk and total$_{tgt}$ to the total number of words of the target chunk):

$$\frac{\text{linked}_{src} + \text{linked}_{tgt}}{\text{total}_{src} + \text{total}_{tgt}} \tag{5.1}$$

The candidate anchor chunk *across the world* ∼ *in de gehele wereld* contains one lexical link (*world* ∼ *wereld*) and two PoS links (*across* ∼ *in* and *the* ∼ *de*). Hence the percentage of linked words = (3 + 3) / (3 + 4) = 0.86.

If a candidate anchor chunk passes the similarity test, the information in the matrix is updated as follows:

- All lexical links inside an anchor chunk are marked with the label **S** (Sure lexical links).
- Words linked based on corresponding part-of-speech codes are marked with the label **p** (PoS links).
- The label **r** is used to mark lexical links that are removed. If a source or target word had multiple lexical links, all lexical links other than the one(s) in the anchor chunk get the label **r**.

5.3.3 Linking adjacent function word chunks

In a final step, chunks consisting of one function word – mostly punctuation marks and conjunctions – can be linked on the basis of corresponding part-of-speech codes if their left or right neighbours on the diagonal are anchor chunks. Corresponding final punctuation marks are also linked.

```
               M| d  V| ,| d  a| o  o  b| d  v  w| h  p| i  N  Y| e| W| h| d  b  v| i  d  g  w| g| e| v| .|
               e| e  o|  | e  a| p  n  u| i  o  e| e  l| n  e  o| n| a| e| e  e  o| n  e  e  e| e| n| e|  |
               v|    o|  |    n|    s  r| e  r  e| b  a|    w  r|  | s| b|    s  l|       h  r| s|  | r|  |
               r|    r|  |    v|    c  g|    i  k| b  a|       k|  | h| b|    c  k|       e  e| c|  | o|  |
               o|    z|  |    a|    h  e|    g   | e  t|        |  | i| e|    h  e|       l  l| h|  | n|  |
               u|    i|  |    l|    u  r|    e   | n  s|        |  | n| n|    a  r|       e  d| o|  | t|  |
               w|    t|  |    l|    l  s|        |    g|        |  | g|  |    a  e|           | k|  | w|  |
                |    t|  |    e|    d   |        |    e|        |  | t|  |    f  n|           | t|  | a|  |
                |    e|  |    n|    i   |        |    v|        |  | o|  |    d   |           |  |  | a|  |
                |    r|  |     |    g   |        |    o|        |  | n|  |    e   |           |  |  | r|  |
                |     |  |     |    e   |        |    n|        |  |  |  |        |           |  |  | d|  |
                |     |  |     |        |        |    d|        |  |  |  |        |           |  |  | i|  |
                |     |  |     |        |        |    e|        |  |  |  |        |           |  |  | g|  |
                |     |  |     |        |        |    n|        |  |  |  |        |           |  |  | d|  |
                |     |  |     |        |        |     |        |  |  |  |        |           |  |  |  |  |
         Madam S|    S|  |     |        |        |     |        |  |  |  |        |           |  |  |  |  |
     President  |    S|  |     |        |        |     |        |  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
             ,  |     | p|     |        |        |     |        |  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
          last  |     |  |     |        |    3   |     |        |  |  |  |        |           |  |  |  |  |
        week's  |     |  |     |        |       1|     |        |  |  |  |        |           |  |  |  |  |
       attacks  |     |  |    4|        |        |     |        |  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
            on  |     |  |     | p      |        |     |        |  |  |  |        |           |  |  |  |  |
      innocent  |     |  |     |    S   |        |     |        |  |  |  |        |           |  |  |  |  |
     civilians  |     |  |     |       S|        |     |        |  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
            in  |     |  |     |        |        |     | p      |  |  |  |        |           |  |  |  |  |
           New  |     |  |     |        |        |     |    S   |  |  |  |        |           |  |  |  |  |
          York  |     |  |     |        |        |     |    S  S|  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
           and  |     |  |     |        |        |     |        | p|  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
    Washington  |     |  |     |        |        |     |        |  | S|  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
       shocked  |     |  |     |        |        |     |        |  |  |  |        |           | S|  |  |  |
-----------------------------------------------------------------------------------------------------------
           and  |     |  |     |        |        |     |        |  |  |  |        |           |  | p|  |  |
-----------------------------------------------------------------------------------------------------------
      outraged  |     |  |     |        |        |     |        |  |  |  |        |           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
           the  |     |  |     |        |        |     |        |  |  |  | p      |           |  |  |  |  |
     civilised  |     |  |     |        |        |     |        |  |  |  |    S   |           |  |  |  |  |
        people  |     |  |     |        |        |     |        |  |  |  |       p|           |  |  |  |  |
-----------------------------------------------------------------------------------------------------------
        across  |     |  |     |        |        |     |        |  |  |  |        | p         |  |  |  |  |
           the  |     |  |     |        |        |     |        |  |  |  |        |    p      |  |  |  |  |
         world  |     |  |     |        |        |     |        |  |  |  |        |          S|  |  |  |  |
-----------------------------------------------------------------------------------------------------------
             .  |     |  |     |        |        |     |        |  |  |  |        |           |  |  |  | p|
-----------------------------------------------------------------------------------------------------------
```

Figure 5.5: Matrix containing anchor chunks (**S** and **p** labels) and remaining lexical links with frequency weights.

The resulting matrix for our example is shown in figure 5.5. In the example, the following anchor chunks were retrieved:

Madam President	Mevrouw | de Voorzitter
,	,
on innocent civilians	op onschuldige burgers
in New York	in New York
and	en
Washington	Washington
shocked	geschokt
and	en
the civilised people	de beschaafde volkeren
across the world	in de gehele wereld
.	

All retrieved anchor chunks but one can be considered to be entirely correct: *shocked* should have been linked to *hebben ... geschokt*, so the anchor chunk *shocked* - *geschokt* is only partially correct. In section 5.4, we describe how we evaluated the performance of the system.

5.4 Experimental results

5.4.1 Training and test data

To evaluate the system's performance, the links created by the system were compared with the links of the manually created reference files. When this work was carried out, only three text types were available: proceedings of European plenary debates (see section 2.2.1), press releases (see section 3.6.3) and user manuals (see section 3.6.6).

From the manually annotated files, 70% were considered as training data, the remaining 30% as test data. Table 5.1 gives an overview of the number of words and documents used for testing the system.

As explained in chapter 4, to be able to compare the alignments of the system with the reference alignments, all phrase-to-phrase alignments were converted into word-to-word alignments by linking each word of the source phrase to each word of the target phrase. All results are presented in terms of weighted F-measure (see section 4.4).

Text type	# Words	# Texts
Proceedings EP	3,139	7
Press releases	4,926	4
User manuals	4,010	2
Total	12,075	13

Table 5.1: English-Dutch test data used in this chapter.

5.4.2 Results

We report on experiments with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries automatically extracted from bilingual corpora of various sizes.

The handcrafted bilingual lexicon was derived from the English-Dutch and Dutch-English NL-Translex lexica (see section 2.4.1). The resulting bilingual dictionary derived from the NL-Translex lexica contains 58,970 English-Dutch word pairs stored as lemmata.

In order to investigate the impact of the size of the training corpus used for dictionary creation and hence the size of the resulting probabilistic dictionary, we extracted bilingual dictionaries from a 2.4 and a 9.3 million word subcorpus of the Europarl corpus (see section 2.4.3). The resulting probabilistic bilingual dictionaries contain respectively 16,728 and 28,342 word pairs stored as word forms.

In the first set of experiments, we looked at the performance of the sub-sentential alignment system on different text types. The sub-sentential alignment system uses the probabilistic dictionary trained on the 9.3M word corpus as lexical resource. We considered all word-to-word links at two different stages in the system. A first time after dictionary look-up, and a second time after the alignment of chunks on the basis of syntactic similarity. However, as it is also interesting to assess the reliability of the alignments of the anchor chunks, precision and recall are also calculated on only the word-to-word links that are part of the aligned anchor chunks.

Table 5.2 contains the results per text type for all the texts of the test corpus.

In the columns under the heading *DCT*, the results after dictionary look-up are displayed. It is worth noticing that although the bilingual lexicon was trained on Europarl data, the coverage is quite good on other domains. The high figure for press releases can be explained by the high number of technical terms that are identical in source and target text (the press releases test corpus contained 12% identical strings).

The columns under the heading *DCT + AC* contain the results after the alignment of syntactically similar chunks. During the alignment process, extra word-to-word links can be created for words belonging to the anchor chunks on the basis of corresponding PoS codes (e.g. function words, adjectives). This explains the increase in recall. On the other hand, some disambiguation takes place for words that were linked several times, which explains the increase in precision.

In the last columns under the heading *AC*, precision and recall scores are only given for the word-to-word links belonging to the chunks that were aligned based on syntactic similarity. The obtained precision scores (between .92 and .98) seem high enough to use the aligned chunks as anchors in the second phase of the alignment process.

	Dct			Dct + Ac			Ac		
	F	Prec	Rec	F	Prec	Rec	F	Prec	Rec
Proceedings EP	.41	.78	.28	.59	.90	.44	.54	.92	.38
Press releases	.49	.90	.34	.74	.98	.59	.70	.98	.54
User manuals	.41	.85	.27	.62	.95	.46	.57	.97	.40

Table 5.2: Weighted F-measure, precision and recall on all word-to-word links at different stages in the system per text type. The system uses the probabilistic dictionary trained on the 9.3M word corpus as lexical resource.

It is also interesting to see how many links generated by the system belonged to either regular or fuzzy alignments in the reference corpus. In table 5.3, the results for the system that uses the probabilistic dictionary trained on the 9.3M word corpus are presented. Table 5.3 shows that the majority of word-to-word links found by the system are regular links. This confirms expectations, as fuzzy links were used for translation-specific shifts of various kinds and are not likely to be captured as anchor chunks.

In the second set of experiments, we used different types of bilingual dictionaries as lexical resource. We investigated the impact of the size of the training corpus used for dictionary creation on the test results. We compared the obtained weighted F-measure, precision and recall scores at the different stages in the system on all the test files. As can be seen in table 5.4, the size of the training corpus – and hence the size of the resulting dictionary – has a positive impact on recall at all stages in the system. No difference in precision was observed.

As explained in section 2.4.3, the probabilistic dictionaries were automatically

	Dct		Dct + Ac		Ac	
	Reg.	Fuz.	Reg.	Fuz.	Reg.	Fuz.
Proceedings EP	.81	.19	.82	.18	.84	.16
Press releases	.94	.06	.95	.05	.97	.03
User manuals	.88	.12	.91	.09	.94	.06

Table 5.3: Percentage of regular and fuzzy links generated by the system per text type. The system uses the probabilistic dictionary trained on the 9.3M word corpus as lexical resource.

extracted from a corpus containing word forms. We examined the impact of lemmatizing the training corpus prior to dictionary creation on the test results. By lemmatizing the training corpus, we expect that abstracting over word forms will increase the overall recall scores.

As the size of the dictionary might influence the impact of lemmatization, we trained probabilistic dictionaries on four different corpora: the lemmatized and word form version of both the 2.4M and the 9.3M word corpus respectively[3].

	Dct			Dct + Ac			Ac		
	F	Prec	Rec	F	Prec	Rec	F	Prec	Rec
2.4M word forms	.43	.85	.29	.64	.95	.48	.60	.97	.43
9.3M word forms	.44	.85	.30	.66	.95	.51	.62	.97	.46
2.4M lemmata	.43	.85	.29	.65	.95	.49	.60	.96	.44
9.3M lemmata	.45	.85	.31	.66	.95	.51	.61	.96	.45
NL-Translex	.42	.87	.28	.65	.96	.49	.61	.97	.44

Table 5.4: Weighted F-measure, precision and recall on all word-to-word links at different stages in systems using different types of bilingual dictionaries.

As can be seen in table 5.4, lemmatization has only a minor positive impact on recall for the system using the 2.4M word corpus. The recall improvement is less clear for the system using the 9.3M word corpus. The reason for this is that the coverage of the word form dictionaries trained on the larger corpora is quite high. We lemmatized the resulting dictionary extracted from the 9.3M word forms corpus off-line and compared it with the dictionary extracted from the lemmatized 9.3M corpus. An overlap of 95% was obtained. The overlap on the resulting dictionaries trained on the 2.3M word corpus was 85%. The precision

[3] In the lemmatized systems, the retrieval of the lemmatized form is not penalized by reducing the frequency weight.

scores are slightly better for the word forms corpora.

It is also interesting to compare the performance of the handcrafted bilingual lexicon with the performance of the probabilistic bilingual lexicon. Although the NL-Translex dictionary is twice the size of the probabilistic dictionary trained on the 9.3M word corpus, the obtained recall scores are a bit lower at all stages of the system. It is also worthwhile mentioning that the NL-Translex dictionary and the probabilistic dictionary contain different word pairs: only 21% of the entries of the probabilistic dictionary are part of the NL-Translex dictionary. The alignments retrieved by the system using the NL-Translex system are a bit more precise after dictionary look-up. But no difference in precision is observed if we only take into account the retrieved anchor chunks.

5.5 Qualitative analysis

In order to evaluate where the remaining problems lie and to assess how the system can be further improved, we inspected the links created at two stages in the system: after the creation of the lexical link matrix and after anchor chunk alignment.

5.5.1 Lexical Link Matrix

We manually inspected the wrong and missed word-to-word alignments after dictionary look-up to reveal the weaknesses of the system.

The precision errors are mainly due to the fact that the bilingual dictionary contains *anomalous* word pairs, such as $I \sim$ *heb*, and $I \sim$ *ben*, which were extracted from the parallel corpus on the basis of co-occurrence frequency values, next to normal entries such as $I \sim$ *ik*. In most cases, the probability values for these *anomalous* entries are lower than the probability values of the normal entries, so most errors are solved in the process of anchor chunk alignment. However, some errors still remain. An extra check on non-compatible word classes, (e.g. verbs cannot be linked to pronouns) might solve this problem.

Most recall errors belong to the following categories: out-of-vocabulary words, compounds, and separable verbs.

- Out-of-vocabulary words.
  The set of out-of-vocabulary words contains word pairs that are not included in the bilingual dictionary. Examples were found for frequent word

pairs, e.g. *window* ∼ *venster* and less frequent or more specific word pairs, e.g. *position* ∼ *betrekking* and *authorization* ∼ *volmacht.*

- Dutch compounds.
  Dutch and English use a different compounding strategy. In English, most compounds are multiwords, whereas in Dutch, compounds are often not separated by means of white space characters and hence form a single word. In some cases, the probabilistic bilingual dictionarydoes not contain the compound as a whole, but one or more parts of the compound are included in the bilingual dictionary. For example *client-independent* ∼ *klantonafhankelijk* is not a dictionary entry, but *client* ∼ *klant* and *independent* ∼ *onafhankelijk* are.

- Separable verbs.
  A specific problem for Dutch are verb particles that can be separated from the verb, e.g. *address* ∼ *aan ... pakken.* The particle can occur before or after the verb and can be separated from the verb by other constituents or clauses. Such particles were not linked by the system. In some cases, the verbal part was linked.

A solution to improve the coverage of the bilingual dictionary is using a larger and more varied bilingual sentence-aligned corpus to extract the bilingual dictionary. However, this solution is not very realistic as large parallel corpora are not available for all domains. To improve the word alignment results for Dutch, the development of a module that can segment compounds into its meaningful parts and a module that can group separable verbs with the verb particle will improve the coverage of the probabilistic dictionary.

5.5.2 Anchor Chunk Alignment

To assess the quality of the aligned anchor chunks, we extracted all aligned chunks that are compatible with the manual word-to-word reference alignments. Four different types of aligned chunks were distinguished:

- One-to-one chunks
  | of asylum seekers | - | van asielzoekers |
  | can be set up | - | kunnen worden aangemaakt |

- One-to-many chunks
  | Mr President | - | Mijnheer | de Voorzitter |
  | on state aids | - | op steunmaatregelen | van overheidswege |

	1:1	1:n	n:1	n:m
Proceedings EP	.68	.11	.04	.17
Press releases	.80	.10	.04	.06
User manuals	.75	.12	.05	.08
Total	.76	.11	.04	.09

Table 5.5: Percentage of one-to-one, one-to-many, many-to-one, and many-to-many chunk alignments in the manual reference alignment per text type.

- Many-to-one chunks
  | every individual | in the audience | - | elke toeschouwer |
  | the heads | of government | - | de regeringshoofden |
- Many-to-many chunks
  | could be | quite drastic | - | kan | ... | immers | escaleren |
  | to make | a statement | - | om een verklaring | uit | te vaardigen |

In the case of one-to-many, many-to-one and many-to-many chunk alignments, the chunks can be discontiguous.

Table 5.5 contains the percentage of one-to-one, one-to-many, many-to-one and many-to-many aligned chunks. As expected, the text type that contained the freest translations – the Europarl data set – contains the highest number of many-to-many chunk alignments, whereas the text types that contain the most literal translations – the press releases and user manuals – contain the highest number on one-to-one aligned chunks.

To asses the quality of the aligned anchor chunks, we calculated precision and recall at the level of aligned chunks. Overall, 55% of the chunks could be linked with a precision of 80%. A more detailed overview is given in table 5.6. More than 95% of the correctly aligned chunks were one-to-one chunks. As explained in chapter 4, F-measure at chunk level is very strict and does not account for partially correct chunk alignments. This explains why the scores calculated at chunk level are much lower than the scores calculated at word level.

5.6 Summary

In this chapter, we have described the global architecture of our sub-sentential alignment system. We conceive our sub-sentential aligner as a cascaded model with two phases.

	F-measure	Precision	Recall
Proceedings EP	.57	.67	.49
Press releases	.74	.85	.65
User manuals	.61	.85	.48
Total	.65	.80	.55

Table 5.6: F-measure, precision and recall calculated at chunk level per text type. The system uses the probabilistic dictionary trained on the 9.3M word corpus as lexical resource.

The objective of the first phase was to link *anchor chunks*, i.e. chunks that can be linked with a very high precision. In our baseline system, on average 45-60% of the words can be linked with a precision ranging from 90% to 98%. If we calculate the more strict F-measure at chunk level, 55% of the chunks could be linked with a precision of 80%. The obtained precision scores seem high enough to use the aligned chunks as anchors in the second phase of the alignment process.

We experimented with two different types of bilingual dictionaries to generate the lexical correspondences: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries. We demonstrated that although the handcrafted dictionary is twice the size of the probabilistic dictionary, the obtained recall scores are lower. No difference in precision is observed for the retrieved anchor chunks.

We demonstrated that lemmatizing the training corpus prior to dictionary extraction can increase recall for small training corpora. As expected, increasing the size of the training corpora has a positive impact on the overall recall scores.

In the next chapter, we focus on the alignment of more complex translational correspondences and present a chunk-driven bootstrapping approach. In chapter 6, we no longer rely on bilingual dictionaries to retrieve the lexical correspondences, but start from the intersected IBM Model 4 word alignments.

CHAPTER 6

A chunk-driven bootstrapping approach

6.1 Introduction

In the previous chapter, we presented the global architecture of our sub-sentential alignment system. We explained that our alignment system is conceived as a cascaded model consisting of two phases. We described in detail the first phase of the alignment process and introduced the notion of anchor chunks – chunks that can be linked with a very high precision on the basis of lexical correspondences and syntactic similarity. In this chapter, the focus is on the second phase of the process, and we will explain how the anchor chunks of the first phase are used to retrieve more complex translational correspondences.

In the previous chapter, two types of bilingual dictionaries were used as lexical resource: a manually created bilingual dictionary and probabilistic bilingual dictionaries derived from the IBM Model One word alignments. In this chapter, the system does not start from a bilingual dictionary but builds further on the output of GIZA++ (Och and Ney 2003), a state-of-the-art tool for statistical word alignment, which implements the IBM models 1–5 and is able to generate better word alignments than IBM Model One[1].

[1] Part of the work described here is published in Macken and Daelemans (2010).

In the context of statistical machine translation, GIZA++ is one of the most widely used word alignment toolkits. GIZA++ implements the IBM models 1–5[2] (Brown et al. 1993) and is used in Moses (Koehn et al. 2007) – an open-source statistical machine translation system – to generate the initial source-to-target and target-to-source word alignments after which a symmetrization heuristic combines the alignments of both translation directions. *Intersecting* the two alignments results in an overall alignment with a higher precision, while taking the *union—textbf* of the alignments results in an overall alignment with a higher recall. The default symmetrization heuristic applied in Moses (*grow-diag-final*) starts from the intersection points and gradually adds alignment points of the union to link unaligned words that neighbour established alignment points. The main problem with the *union* and the *grow-diag-final* heuristics is that the gain in recall causes a substantial loss in precision, which poses a problem for applications intended for human users.

As can be seen in table 6.1, the intersected GIZA++ alignment points are very precise and have a much higher recall[3] than the alignments resulting from the dictionary lookup in the dictionary derived from the IBM Model One word alignments.

Therefore, in this chapter, we will build our anchor chunks on the basis of the intersected GIZA++ word alignments. The anchor chunks and the intersected word alignments are then used to bootstrap the extraction of more complex translational correspondences (e.g. deletion of a determiner in a noun phrase, change from premodification to postmodification in a noun phrase).

	WF1	WPrec	WRec
Model One Dict.	44.7	85.2	30.3
Intersected GIZA++	74.1	97.2	59.9

Table 6.1: Weighted F-measure, precision and recall on all word-to-word links of a system using a Model One dictionary versus the intersected GIZA++ alignments, calculated on all test files presented in table 5.1.

[2] IBM Model one is a pure lexical model: it only takes into account the word frequencies of the source and target sentences. The higher numbered IBM Models are more complex and take into account word order (distortion) and model the probability that one source word aligns to more than one target word (fertility).

[3] In contrast with the anchor chunk system described in chapter 5, GIZA++ also aligns function words.

Two other changes have been made to the system described in chapter 5: prepositions are considered as separate chunks and a stricter policy was applied for the identification of anchor chunks:

1. Prepositional phrases are used for a wide range of syntactic and semantic functions, most commonly modification and complementation. In the case of a complement, the preposition and the word it complements often function as one unit. For example, the verb *to dispose* and the preposition *of* form one unit with the meaning *to discard*. As our system offers the flexibility to merge smaller chunks, but cannot split chunks, we consider each preposition as a separate chunk that can be grouped either with the verb phrase, noun phrase or adjectival phrase it complements or with the noun phrase it introduces.

2. In the system described in chapter 5, a similarity test was performed on all candidate anchor chunks (see section 5.3.2). In the case of contiguous chunks 80% of the word had to be linked, while in the case of non-contiguous chunks all words had to be linked. In this chapter we apply a stricter policy: only the chunks where all words are linked are considered to be anchor chunks. The reason for this change is that we want to extract the translation-specific patterns (e.g. insertion of a determiner in a noun phrase as in *education* ∼ *het onderwijs* [En: *the education*]) automatically in the bootstrapping process.

In order to assess the impact of those two changes, we compared the results after anchor chunk alignment on the basis of the intersected GIZA++ alignments in two different settings. In the first version, a similarity test threshold of 80% was used in the case of contiguous chunks and prepositions were not considered as separate phrases. In the second version, the similarity test threshold was set to 100% and prepositions were considered as separate phrases. The results are presented in table 6.2. As expected, the results of the second version are a bit more precise, but have a lower recall.

	WF1	WPrec	WRec
Similarity threshold 80%	75.2	96.8	61.5
Similarity threshold 100% + prepositions as separate phrases	74.7	97.2	60.7

Table 6.2: Weighted F-measure, precision and recall on all word-to-word links after anchor chunk alignment on the basis of the intersected GIZA++ alignments calculated on all test files presented in table 5.1.

6.2 Bootstrapping approach

The starting point of the bootstrapping process is a sentence-aligned parallel corpus (see section 6.3.1) in which anchor chunks have been aligned on the basis of the intersected GIZA++ alignments and syntactic similarity.

The bootstrapping process is a cyclic process that alternates between extracting candidate translation rules (extraction step) and scoring and filtering the extracted candidate translation rules (validation step). From the second bootstrapping cycle onwards, the validated translation rules are first applied to the corpus, after which the extraction process is launched again. The bootstrapping process is repeated several times.

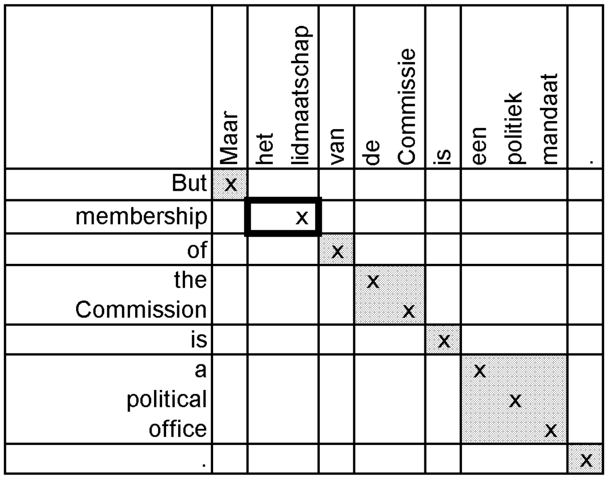
6.2.1 Extraction step

In the extraction step, candidate translation rules are extracted from unlinked source and target chunks. Different alignment types (1:1, 1:n, n:1 and n:m) are considered:

- From sentence pairs that only contain 1:1, 1:n and n:1 unlinked chunks, candidate translation rules that link 1:1, 1:n and n:1 chunks are extracted. In figure 6.1, the source chunk *membership* and target chunk *het lidmaatschap* are selected because they are the only unlinked chunks in the sentence pair.

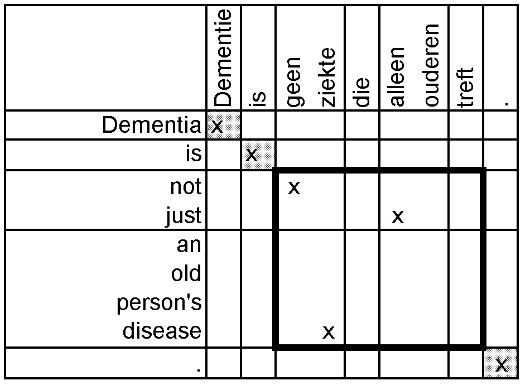
- From sentence pairs in which the only unlinked chunks in the source or target sentence are lexically interlinked, candidate translation rules that link n:m chunks are extracted. In figure 6.2, the source chunks *not just | an old person's disease* and the target chunks *geen ziekte | die | alleen ouderen |treft* [En literal translation: *no disease that just elderly strikes*] are selected, as the selected source chunks are connected by means of the lexical links between *not* and *geen* and *disease* and *ziekte* and are the only unlinked chunks in the source sentence.

From the selected source and target chunks two types of rules are extracted: *abstract* rules and *lexicalized* rules. The rules can be contiguous or non-contiguous.



	Maar	het	lidmaatschap	van	de	Commissie	is	een	politiek	mandaat	.
But	x										
membership			x								
of				x							
the					x						
Commission						x					
is							x				
a								x			
political									x		
office										x	
.											x

Figure 6.1: Sentence pair with one unlinked source (*membership*) and target chunk (*het lidmaatschap* [En: *the membership*]). Chunk boundaries are indicated by horizontal and vertical lines, intersected GIZA++ word alignments by x's, and anchor chunks in light grey.



	Dementie	is	geen	ziekte	die	alleen	ouderen	treft	.
Dementia	x								
is		x							
not			x						
just						x			
an									
old									
person's									
disease				x					
.									x

Figure 6.2: Sentence pair with unlinked source chunks that are grouped by means of lexical links. Chunk boundaries are indicated by horizontal and vertical lines, intersected GIZA++ word alignments by x's, and anchor chunks in light grey.

- Abstract rules are coded as PoS sequences. Established word alignments within the extracted chunks are coded as indices. For example the rule DET+N-gen+N_1 → DET+N_1|PREP|DET+N captures the transformation of a genitive into a prepositional phrase as in *the public's right* → *het recht van de burgers* [En: *the right of the public*].
- Lexicalized rules are coded as token sequences, e.g. *to treat* → *ter behandeling van* [En: *for the treatment of*].

From the second bootstrapping cycle onwards, the validated rules are first applied to the whole training corpus, resulting in new translation pairs containing 1:1, 1:n and n:1 unlinked chunks, after which the extraction process is launched again.

The matching process considers all lexically interlinked groups of chunks (see figure 6.2) and all unlinked source and target chunks with a neighbouring left or right anchor chunk and uses the word aligments and the anchor chunks to build up the target or source pattern.

6.2.2 Validation step

The aim of the validation step is to extract a subset of *reliable* translation rules from the set of candidate translation rules.

We experimented with three different validation methods:

1. Applying the candidate rules to a manually created reference corpus and scoring the candidate translation rules.
2. Filtering the candidate rules on the basis of a number of heuristics.
3. Filtering the candidate rules on the basis of statistical association measures.

Using a reference corpus

In the first method, the set of candidate translation rules is validated against a manually created reference corpus. The extracted translation rules are applied to the validation corpus in order to retain a subset of *reliable* translation rules. The processing steps are visualized in figure 6.3.

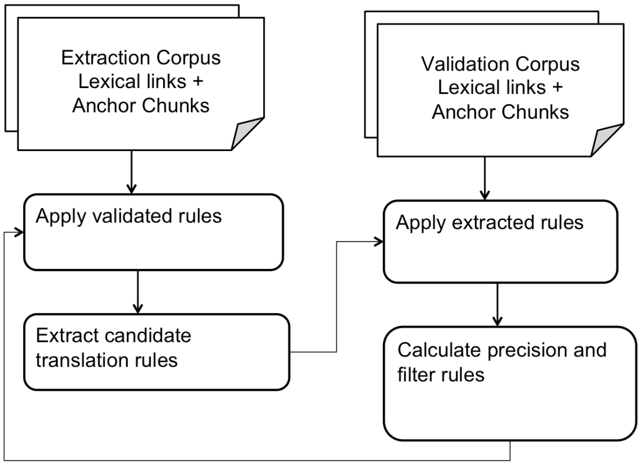


Figure 6.3: Overview of the bootstrapping process using a validation corpus.

We use the chunk information available from the rule-based chunkers in the system and the manually created word links to group together all source and target chunks that are translations of each other. A rule is considered to be correct if it matches a reference linked chunk completely, otherwise it is considered to be incorrect. We illustrate how the rules are scored by means of the following example sentence:

(16) En: The audience in the Shrine Auditorium and the television audience will see the best images possible , ' said Seligman .

Nl: Het publiek in het Shrine Auditorium en het publiek thuis zullen de best mogelijke beelden zien ' , zegt Seligman .

Table 6.3 contains the manually aligned chunks for the example sentence.

Source chunk(s)	Target chunk(s)
The audience	Het publiek
in	in
the Shrine Auditorium	het Shrine Auditorium
and	en
,	,
the television audience	het publiek | thuis
will see	zullen ... zien
the best images possible	de best mogelijke beelden
,	,
said	zegt
Seligman	Seligman
.	.

Table 6.3: Aligned chunks in the manual reference.

Two abstract candidate translation rules were validated in this sentence pair:

- Rule DET_1+N+N_2 → DET_1+N_2 aimed to link the chunk *the television audience* to *het publiek* [En: *the public*]. As these chunks only partially correspond with the aligned chunks in the manually created reference, the rule is considered to be incorrect in this sentence pair.
- Rule V-fin → V-fin aimed to link the chunk *said* to *zegt* [En: *says*]. As these chunks correspond with the aligned chunks in the manually created reference, the rule is considered to be correct in this sentence pair.

The precision of a rule is the number of times the rule was applied correctly divided by the total number of times the rule was applied.

As the precision of a rule is calculated at the level of aligned chunks, it is very strict and does not account for partial correctness (see also section 4.5). Only the rules that were validated with a precision of 80% were retained. This value was set experimentally.

Using filtering heuristics

The two main disadvantages of using a validation corpus is that a manual annotation effort is required and that the number of translation patterns that can be validated is limited to the patterns that occur in the validation corpus.

Therefore, we tried to use filtering heuristics to retain reliable rules. The processing steps are visualized in figure 6.4. We manually inspected the list of rules (sorted according to frequency) and implemented the most straightforward heuristics.

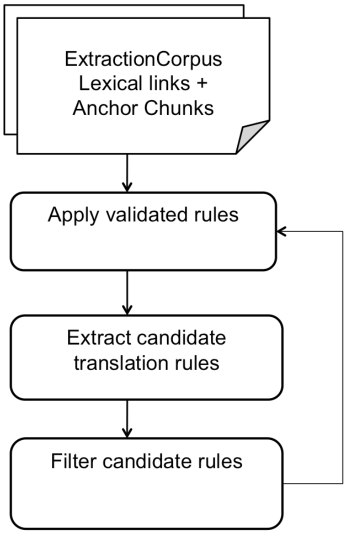


Figure 6.4: Overview of the bootstrapping process using filtering heuristics.

Two heuristics remove rules from the set of candidate rules:

- Rules with an empty source or target pattern are removed.
- Rules in which only the source or target pattern contains a verb form that is not linked to a target or source word by means of a word alignment, e.g. DET | V-fin → PRON-per (*that* | *is* → *we*).

The following heuristics retain translation rules as being reliable:

- Rules for which the source and target pattern form identical phrase patterns (e.g. NP → NP, VP → VP)
- Rules with discontinuous patterns if each contiguous part of the discontinuous rule contains lexical links.
- Rules that were extracted at least 10 times.

Using statistical association measures

The main disadvantage of the validation method using heuristics is that the heuristics are manually created, and therefore somewhat arbitrary. We try to overcome this problem by using statistical association measures that are computed on the extraction corpus. It is a fully automatic method that does not make use of a manual reference corpus.

This last validation method has an additional advantage – apart from being fully automatic: the statistical association scores can also be used to determine the order in which the rules are applied.

The processing steps of this last method are visualized in figure 6.5.

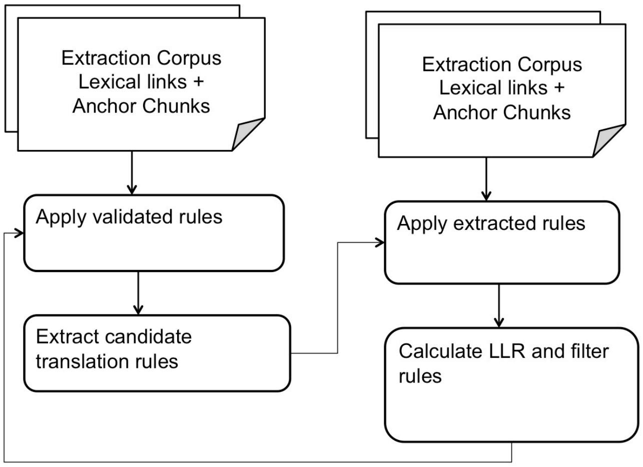


Figure 6.5: Overview of the bootstrapping process using statistical association metrics.

We use the Log-Likelihood ratio (Dunning 1993) as statistical association measure to compute an association score between each source and target pattern of all candidate translation rules. The Log-Likelihood ratio has been used before for building translation dictionaries (Melamed 2000) and for word alignment (Moore 2005).

To compute the Log-Likelihood ratio, we first count for each candidate translation rule how many times the source and target pattern co-occur and store these data in a contingency table containing the observed frequencies (see table 6.4).

In general, table 6.4 can be interpreted as follows:

- O_{11} is the number of times that the source pattern and the target pattern co-occur.
- O_{12} is the number of times that the source pattern occurs and the target pattern does not occur.
- O_{21} is the number of times that the target pattern occurs and the source pattern does not occur.
- O_{22} is the number of times that neither the source pattern nor the target pattern occur.

	tgt	¬ tgt	
src	O_{11}	O_{12}	R_1
¬ src	O_{21}	O_{22}	R_2
	C_1	C_2	N

Table 6.4: Contingency table containing the observed frequencies.

As the Log-Likelihood ratio measures the difference between the observed and expected values under the null hypothesis of independence, we calculate expected frequencies from the row and column totals as shown in table 6.5.

	tgt	¬ tgt
src	$E_{11} = \frac{R_1C_1}{N}$	$E_{12} = \frac{R_1C_2}{N}$
¬ src	$E_{21} = \frac{R_2C_1}{N}$	$E_{22} = \frac{R_2C_2}{N}$

Table 6.5: Expected frequencies for the data in the contingency table.

The Log-Likelihood ratio on the basis of the observed frequencies and the expected frequencies under the null hypothesis of independence is then calculated as follows:

$$-2log(\lambda) = 2\sum{}_{ij}O_{ij}log(\frac{O_{ij}}{E_{ij}}). \tag{6.1}$$

Dunning's Log-Likelihood Ratio gives an indication of the degree of association between a particular source and target pattern. Generally speaking, the higher the Log-Likelihood value, the more significant the difference is between the two frequency scores.

Dunning showed that the Log-Likelihood ratio test allows comparisons to be made between the significance of the occurrences of both rare and common phenomena, which makes the test appropriate for our purposes. According to Manning and Schütze (2003), $-2log(\lambda)$ has a distribution similar to that of chi-square and can therefore be used for hypothesis testing using the statistical tables for the distribution of chi-square. For a contingency table with two rows and two columns the critical value is 10.83 for the significance level of 0.001 (McEnery, Richard and Tono 2006).

Therefore, in the validation step, we only retain translation rules with a Log-Likelihood value higher than 10.8. To reduce the memory requirements of our system, we only validated candidate translation rules that co-occurred at least 5 times.

6.3 Experimental set-up

Different experiments were set up to cover all research parameters:

- Two corpora were used (containing short sentences and medium-length sentences) to extract the candidate translation patterns.
- The three different validation methods described above were applied.

6.3.1 Extraction Corpus

For the extraction step of the bootstrapping process, we created two subcorpora from the Dutch Parallel Corpus (see section 2.2.3) containing sentences of different length.

- The first subcorpus contains 36,406 sentence pairs (478,002 words) of short sentences (1-10 words).
- The second subcorpus contains 79,814 sentence pairs (1,892,233 words) of medium-length sentences (1-20 words).

We opted for subcorpora containing short and medium-length sentences, as short and medium-length sentences are more likely to contain only few unlinked chunks.

The Dutch Parallel Corpus has a balanced composition and contains five text types: administrative texts, texts treating external communication, literary texts, journalistic texts and instructive texts. All text types are present in the selected subcorpora.

6.3.2 Validation Corpus

Two validation methods make use of a validation corpus. In the first validation method, the set of candidate translation rules is validated against a manually created reference corpus. We used the training data section of the manually created reference corpus (described in section 5.4.1) as validation corpus.

The validation corpus is a subcorpus of the manually created reference corpus (described in section 3.6) and contains 24,663 words of three different text types (proceedings of plenary debates, press releases and user manuals).

In the third validation method (*using statistical association metrics*), no manually created reference corpus is needed. In this validation method, the extraction corpus is also used in the validation step.

The second validation method does not make use of a validation corpus. It makes use of manually created filtering heuristics to retain reliable translation patterns.

6.3.3 Test Corpus

To asses the performance of the different bootstrapping systems, the links created by the different systems are compared with the links of the manual reference files. Table 6.6 gives an overview of the number of words, sentences and documents used for testing the different systems. The test data consists of the test set that was used in chapter 5 (see section 5.4.1) along with additional test material extracted from the Dutch Parallel Corpus. None of the test files are part of the extraction or validation corpus.

To be able to compare the alignments of the system with the reference alignments, all phrase-to-phrase alignments were converted into word-to-word alignments by linking each word of the source phrase to each word of the target phrase (all-pairs heuristic).

Text type	# Words	# Sentences	# Texts	Avg. sent. length
Journalistic texts	8,557	354	3	24.2
Proceedings EP	3,139	105	7	29.9
Newsletters	12,000	688	2	17.4
Press releases	4,926	212	4	23.2
Technical texts	8,661	432	4	20.0
User manuals	4,010	296	2	13.5
Total	41,293	2,087	22	19.8

Table 6.6: Characteristics of the English-Dutch test data.

6.4 Results

6.4.1 Impact of each bootstrapping cycle

To get an idea of the impact of each bootstrapping cycle, we logged the percentage of unaligned chunks in the extraction corpus, the number of extracted candidate rules and the total number of validated rules after each bootstrapping cycle in three different bootstrapping systems. The three systems retrieve their translation rules from the extraction corpus of short sentences, but use a different validation method. After each bootstrapping cycle, we applied the set of validated rules on all files of the test corpus and calculated weighted F-measure. The results are presented in table 6.7.

In each bootstrapping cycle, we expect, on the one hand, that the system validates new rules, and, on the other hand, that the percentage of unlinked chunks in the extraction corpus decreases. We observe this expected behaviour in all three systems. However, the three systems differ considerably with respect to the number of validated rules and as a result the number of unlinked chunks in the extraction corpus.

As expected, the validation method using a manually created reference corpus has the lowest total number of validated rules (222 versus 2,274 and 2,354) as the system can only validate patterns that are present in the validation corpus. The last column of table 6.7 displays the number of rules that were actually found in the validation corpus.

The systems using filtering heuristics or Log-Likelihood ratio as validation method show a similar behaviour with regard to learning rules. In the first two bootstrapping cycles, 90% of all validated rules are retrieved. In the fourth bootstrapping cycle, very few new rules are learnt. The two systems are comparable

Validation Corpus

	Unlinked chunks	Extracted patterns	Validated patterns	WF1	Tested patterns
Iteration 1	0.39	14,278	179	75.5	560
Iteration 2	0.34	14,509	206	75.7	179
Iteration 3	0.33	14,547	209	75.7	209
Iteration 4	0.32	14,600	214	75.7	214

Filtering heuristics

	Unlinked chunks	Extracted patterns	Validated patterns	WF1	
Iteration 1	0.39	14,278	1,715	76.5	
Iteration 2	0.25	12,064	2,209	76.5	
Iteration 3	0.23	11,368	2,270	76.6	
Iteration 4	0.23	11,289	2,276	76.6	

Log-Likelihood

	Unlinked chunks	Extracted patterns	Validated patterns	WF1	
Iteration 1	0.39	14,278	1,552	76.7	
Iteration 2	0.22	14,818	2,148	76.8	
Iteration 3	0.20	14,814	2,307	76.8	
Iteration 4	0.20	14,705	2,354	76.8	

Table 6.7: Percentage of unlinked chunks in the extraction corpus, the number of extracted candidate patterns, the total number of validated patterns after each bootstrapping cycle for three different validation methods and the results expressed in terms of weighted F-measure on all test files described in table 6.6.

with respect to the total number of validated rules, but to a large extent, they extract different rules: only 20% of the rules overlap.

The weighted F-measure of the intersected GIZA++ word alignments on all test files was 73.8; after anchor chunk alignment it was 74.7. In all validation methods, the largest improvement in weighted F-measure could be observed after the first iteration. To reduce the memory requirements of our system, we used four bootstrapping cycles in all experiments that we discuss below.

6.4.2 Comparison with symmetrized GIZA++ alignments

We also compared the performance of different bootstrapping systems with the commonly used symmetrized GIZA++ predictions (intersection, union and grow-diag-final) on six different text types.

The results of all our experiments are summarized in tables 6.8 and 6.9 and are expressed in terms of Alignment Error Rate (see 4.3) and weighted F-measure (see 4.4).

In table 6.8, we give per text type the alignment scores for the symmetrized GIZA++ predictions, using the three most commonly used symmetrization heuristics: intersection (∩), union (∪), and grow-diag-final (Gdf).

As already mentioned in section 6.1, GIZA++ is a state-of-the-art tool for statistical word alignment that implements IBM Models 1-5 and is one of the most widely used word alignment toolkits in the context of statistical machine translation.

When we compare the three symmetrization heuristics – as expected – the intersection heuristic generates the most precise overall alignment, while the union heuristic results in an alignment with the highest recall. The recall gain in the *Union* and *Grow-diag-final* heuristics causes a substantial loss in precision.

The results also reflect the different translation strategies of the different text types: the technical texts are the easiest to align; the journalistic and Europarl texts the most difficult. The results for the union and grow-diag-final heuristics are remarkably worse for the journalistic texts than for other text types. Best overall results – expressed both in terms of AER and in terms of weighted F-measure – are obtained by the grow-diag-final heuristic (the default heuristic used in Moses) for all text types, except for journalistic texts.

In table 6.9, the results of our chunk-based extension to the intersected GIZA++ alignments are given for four different settings using the extraction corpus of

	Journalistic			Europarl		
	∩	∪	GDF	∩	∪	GDF
Prec	95.7	57.9	62.0	94.1	73.7	76.1
Rec	65.5	84.2	83.7	64.0	80.0	79.1
AER	21.8	32.4	29.6	22.9	23.6	22.6
WPrec	95.7	58.8	62.9	94.1	75.7	77.9
WRec	51.5	67.6	67.0	51.1	64.9	63.9
WF1	67.0	62.9	64.9	66.2	69.9	70.2

	Newsletters			Press releases		
	∩	∪	GDF	∩	∪	GDF
Prec	96.4	72.3	76.3	98.6	76.2	80.7
Rec	65.4	84.7	83.9	63.3	76.3	75.5
AER	21.8	22.3	20.3	22.7	23.8	21.9
WPrec	96.4	72.5	76.5	98.6	77.3	81.7
Wrec	58.9	75.5	74.7	64.4	76.3	75.8
WF1	73.1	74.0	75.6	77.9	76.8	78.6

	Technical			User manuals		
	∩	∪	GDF	∩	∪	GDF
Prec	97.8	78.0	81.3	97.8	73.2	77.8
Rec	73.2	88.0	87.4	64.1	83.4	82.5
AER	16.1	17.5	15.9	22.3	22.3	20.0
WPrec	97.8	78.5	81.8	97.8	74.1	78.7
WRec	68.1	80.9	80.3	61.0	78.1	77.5
WF1	80.3	79.7	81.1	75.1	76.1	78.1

Table 6.8: Results for the different symmetrized GIZA++ predictions: intersection (∩), union (∪), and grow-diag-final (Gdf) expressed in terms of AER and weighted F-measure.

short sentences:

- **10Cor**: extraction corpus of short sentences (1-10 words), rules validated against a manually created reference corpus.
- **10Filt**: extraction corpus of short sentences (1-10 words), rules validated via filtering heuristics.
- **10All**: extraction corpus of short sentences (1-10 words); Log-Likelihood ratio as validation method.
- **10Lex**: extraction corpus of short sentences (1-10 words); Log-Likelihood ratio as validation method; only lexicalized translation rules or abstract rules containing lexical indices are retained in the validation step.

When we compare the three validation methods, in general, the best results are obtained by the system that uses the Log-Likelihood ratio to filter all rules (*10All).* When only lexicalized translation rules or abstract rules containing lexical indices are retained in the validation step, a small improvement in precision can be observed. In general, the validation method using the manually created reference corpus shows the highest precision scores.

In all settings, the results show an overall AER reduction over all symmetrized GIZA++ predictions. In terms of weighted F-measure, the best performing bootstrapping system (*10All*) showed a higher F-measure for all text types except for the Europarl texts.

	Journalistic				Europarl			
	10Cor	10Filt	10All	10Lex	10Cor	10Filt	10All	10Lex
Prec	94.9	93.1	93.0	94.0	93.9	93.0	92.2	93.8
Rec	70.4	71.3	71.8	70.1	67.3	68.4	68.3	67.8
AER	18.8	18.9	18.6	19.3	20.7	20.3	20.7	20.4
WPrec	94.9	93.2	93.1	94.1	93.8	93.4	92.5	93.9
WRec	55.5	56.2	56.4	55.0	53.8	54.6	54.5	54.1
WF1	70.0	70.1	70.3	69.4	68.4	68.9	68.6	68.6

	Newsletters				Press releases			
	10Cor	10Filt	10All	10Lex	10Cor	10Filt	10All	10Lex
Prec	96.1	95.1	94.9	96.0	98.2	97.8	97.5	98.2
Rec	69.5	71.3	72.0	69.8	65.7	66.2	66.4	65.3
AER	19.1	18.2	17.8	18.9	21.0	20.8	20.8	21.3
WPrec	94.9	95.1	94.9	95.9	98.2	97.9	97.7	98.3
WRec	55.5	63.9	64.4	62.4	66.9	67.2	67.7	66.5
WF1	70.0	76.4	76.7	75.7	79.6	79.9	79.9	79.3

	Technical				User manuals			
	10Cor	10Filt	10All	10Lex	10Cor	10Filt	10All	10Lex
Prec	97.6	96.6	96.3	97.2	97.1	96.5	96.4	96.6
Rec	76.3	77.3	77.6	76.3	68.0	69.4	70.0	68.0
AER	14.2	13.9	13.8	14.3	19.7	19.0	18.5	19.8
WPrec	97.6	96.7	96.4	97.3	97.1	96.5	96.4	96.7
WRec	71.0	71.9	72.3	70.9	64.8	65.8	66.8	65.0
WF1	82.2	82.4	82.6	82.0	77.7	78.3	78.9	77.7

Table 6.9: Results for the chunk-based extension to the intersected GIZA++ alignments for four different settings expressed in terms of AER and weighted F-measure.

We tried to improve the overall results of our bootstrapping system that uses the Log-Likelihood ratio as validation method in two ways:

1. From the results presented in table 6.9, we can deduce that the abstract rules without lexical clues are the least precise, but improve the recall of the system. We therefore added one extra bootstrapping cycle to the *10Lex* system that retains only lexicalized translation rules or abstract rules containing lexical indices in the validation step. In this extra bootstrapping cycle abstract rules without lexical clues are extracted and validated. All lexicalized translation rules or abstract rules containing lexical indices are applied first; in a second step the abstract rules without lexical clues are applied.
2. We used the larger extraction corpus of medium-length sentences.

We combined the two ways of improvement and set up experiments with the following three systems:

- **10LexAbs**: identical to 10Lex but with one additional bootstrapping cycle to extract and validate abstract rules without lexical clues. All lexicalized translation rules or abstract rules containing lexical indices are applied first; in a second step the abstract rules without lexical clues are applied.
- **20Lex**: identical to 10Lex but using an extraction corpus of medium-length sentences (1-20 words).
- **20LexAbs**: identical to 10LexAbs but using an extraction corpus of medium-length sentences (1-20 words).

The best results are obtained by the system that uses the larger extraction corpus and that first extracts and applies rules with lexical clues and in a second pass the more abstract rules without lexical clues (*20LexAbs*). In general, better results are obtained by those 2-pass systems. It seems that these systems are better able to control the problem of error propagation. Enlarging the extraction corpus had a positive effect on the 2-pass systems, but the effect is less clear on the systems that only allow lexicalized rules or rules with lexical clues.

	Journalistic			Europarl		
	10LexAbs	20Lex	20LexAbs	10LexAbs	20Lex	20LexAbs
Prec	93.0	92.4	92.0	92.4	93.4	92.0
Rec	71.9	70.6	73.1	68.4	68.7	69.2
AER	18.6	19.7	**18.2**	20.5	**19.9**	20.1
WPrec	93.1	93.3	92.1	92.8	93.5	92.4
WRec	56.6	55.7	57.6	54.8	54.8	55.4
WF1	70.4	69.8	**70.9**	68.9	**69.1**	69.3

	Newsletters			Press releases		
	10LexAbs	20Lex	20LexAbs	10LexAbs	20Lex	20LexAbs
Prec	94.9	94.6	94.6	97.6	97.7	96.8
Rec	71.8	70.2	72.4	66.4	65.8	66.9
AER	18.0	19.1	**17.7**	20.7	21.1	**20.6**
WPrec	94.9	95.8	94.6	97.8	97.7	97.0
WRec	64.3	63.1	64.8	67.6	67.0	68.1
WF1	76.7	76.1	**76.9**	79.9	79.5	**80.1**

	Technical			User manuals		
	10LexAbs	20Lex	20LexAbs	10LexAbs	20Lex	20LexAbs
Prec	96.4	96.1	96.3	96.4	96.3	95.7
Rec	77.7	77.3	78.4	69.9	68.8	70.8
AER	13.7	14.1	**13.3**	18.6	19.4	**18.3**
WPrec	96.6	96.9	96.4	96.5	96.4	95.8
WRec	72.3	72.1	73.0	66.6	65.7	67.2
WF1	82.7	82.6	**83.1**	78.8	78.1	**79.0**

Table 6.10: Results for the chunk-based extension to the intersected GIZA++ alignments for three different settings expressed in terms of AER and weighted F-measure. The best overall results are indicated in bold.

6.4.3 Summary of the experiments

Table 6.11 gives an overview of the results of the most important systems on all test files. The intersected GIZA++ alignments were taken as a starting point. The intersection heuristic generated the most precise alignment, while the union and the grow-diag-final heuristics generate alignments with the highest recall at the cost of a reduced precision.

The different chunk-based extensions to the intersected GIZA++ alignments improve the recall of the intersected GIZA++ word alignments without sacrificing precision. If we compare the three different validation methods using *a reference corpus* (10Cor), *filtering heuristics* (10Filt), and *Log-Likelihood ratio* (10LexAbs), *10LexAbs* generates the best overall results, and *10Cor* is the most precise system. Enlarging the extraction corpus has a positive effect on the 2-pass system using Log-Likelihood ratio as validation method.

	GIZA++			Chunk-based extension			
	∩	∪	GDF	10Cor	10Filt	10LexAbs	20LexAbs
Prec	96.9	70.9	74.8	96.4	95.4	95.2	94.7
Rec	66.5	83.6	82.9	70.2	71.4	71.8	72.6
AER	20.8	23.6	21.6	18.4	18.0	17.8	**17.5**
WPrec	96.9	71.6	75.5	96.5	95.5	95.3	94.8
WRec	59.6	74.6	73.9	62.2	63.9	64.3	65.0
WF1	73.8	73.1	74.7	75.7	76.6	76.8	**77.1**

Table 6.11: Results for the different GIZA++ predictions and four different settings of the chunk-based extensions to the intersected GIZA++ alignments on all test files described in table 6.6. The results are expressed in terms of AER and weighted F-measure. The best results are indicated in bold.

6.4.4 F-measure calculated at chunk level

To asses the quality of the aligned anchor chunks, we calculated precision and recall at the level of aligned chunks. The results for the symmetrized GIZA++ alignments, and different bootstrapping systems are presented in table 6.12. As explained in chapter 4, F-measure at chunk level is very strict and does not account for partially correct chunk alignments.

The intersected GIZA++ alignments give the most precise results with a precision score of 90.2[4], but is not suited to capture other chunk types than one-to-one chunks. The best overall results are obtained by the *10Filt* system.

	F-measure	Precision	Recall
∩	71.3	90.2	58.9
∪	70.4	81.6	61.8
GDF	70.5	81.6	62.1
10Cor	73.4	87.2	63.4
10Filt	74.4	83.1	67.4
10LexAbs	74.1	84.5	66.0
20LexAbs	74.3	82.2	67.7

Table 6.12: F-measure, precision and recall calculated at chunk level for different systems on all test files described in table 6.6.

6.4.5 Types of rules

Table 6.13 gives an overview of the total number of validated rules in the different experimental settings and gives details on the number of discontiguous (either abstract or lexicalized) and lexicalized validated rules. As expected, the number of validated rules increases when the corpus size is increased. If only translation rules that contain lexical clues are allowed, the number of validated translation rules is drastically reduced. The share of discontiguous rules ranges from 23 to 39%; the share of lexicalized rules from 31 to 48%.

	Validated			Applied		
	Total	Discont.	Lexical.	Total	Discont.	Lexical.
10Lex	1526	344	724	303	46	70
10LexAbs	2135	574	744	508	104	70
20Lex	3790	1174	1828	530	108	153
20LexAbs	5826	2287	1828	872	249	153

Table 6.13: Total number of validated rules in the different settings and number of validated discontiguous and lexicalized rules; total number of applied rules in the test corpora and number of applied discontiguous and lexicalized rules.

On the right-hand side of the table, the number of applied rules in the different test corpora is given. In order to process the 40,000 words of the test corpora,

[4] It is not possible to compare the results with the results obtained in chapter 5 as the rule-based chunker was adapted and considers prepositions as separate chunks in this chapter.

14 to 24% of the rules are applied. The share of discontiguous rules accounts for 14 to 20%.

Some example rules are given below:

- N_1 → DET+N_1 (*History → de Geschiedenis*)
- DET_1+N_2+N → DET_1+N_2 (*a movie producer → een filmproducent*)
- PREP_1+V-prpa_2 → PREP_1 | DET+N_2 | PREP (*for managing → voor het management van*)
- DET_1+N_2;PREP | N_3 → DET_1+N_2+N_3 (*a number of events → een aantal evenementen*)
- V-fin_1+V-papa_2 → V-fin_1 ... V-papa_2 (*had written → had ... geschreven*)
- ADV_1;...;ADJ_2 → ADV_1+ADJ_2 (*not;...;longer → niet langer*)
- last → voor het laatst
- has → beschikt ... over
- agree → ben | het ... eens

The most frequently applied rules take care of the insertion of a determiner in Dutch (e.g. *History → de Geschiedenis*) or deal with Dutch compounds of which only a part has been aligned by the GIZA++ intersected word alignments (e.g. *filmproducent*). The most frequently applied discontiguous rules deal with verbal groups that are often split in Dutch (*had ... geschreven*). However, other discontiguous chunks are captured as well. Some discontiguous lexicalized rules are able to deal with phrasal verbs (e.g. *beschikken ... over*).

An error analysis of the alignments of the proposed system reveals that some structures are only partially linked. Grouping words together before chunk alignments (e.g. Dutch separable verbs – which are often discontinuous, and English phrasal verbs) before chunk alignment may improve the system. Another remaining problem that is only partially solved by our approach is the problem of Dutch compounds. Adding a decompounding module for Dutch prior to GIZA++ alignment might offer a solution to this problem.

6.5 Summary

We developed a new chunk-based method to add language-pair specific knowledge – derived from shallow linguistic processing tools – to statistical word alignment models. The system is conceived as a cascaded model consisting of two phases. In the first phase *anchor chunks* are linked on the basis of the intersected GIZA++ word alignment and syntactic similarity. In the second phase, we use a bootstrapping approach to extract language-pair specific translation patterns.

We developed three different methods to validate the extracted candidate rules in the bootstrapping process. The first method makes use of a manually created reference corpus. The second uses manually created filtering heuristics. The third is fully automatic and filters on the basis of the Log-Likelihood ratio. It was demonstrated that the fully automatic system achieves good results.

We demonstrated that the proposed systems improve the recall of the intersected GIZA++ word alignments without sacrificing precision, which makes the resulting alignments more useful for incorporation in CAT-tools or bilingual terminology extraction tools. Moreover, the system's ability to align discontiguous chunks makes the system useful for languages containing split verbal constructions.

In the next chapters, we move to applications. In chapter 7, we describe how the anchor chunk alignment module was used to guide bilingual terminology extraction. In this chapter we also demonstrate the language independence of our approach as the terminology extraction experiments were carried out on three different language pairs, viz. French-English, French-Italian and French-Dutch.

In chapter 8, we will demonstrate that sub-sentential alignment is an important component of sub-sentential translation memory systems.

CHAPTER 7

Bilingual terminology extraction

7.1 Introduction

Terminology can be defined as the study of terms and encompasses diverse activities such as the collection, description and structuring of terms. According to Wright (1997, p. 13), terms are “the words that are assigned to concepts used in the special languages that occur in subject-field or domain-related texts”.

There is an undeniable link between terminology and technical translation, which is why most specialized translation training programs include terminology courses in their curriculum. Bowker (2008, p. 286) states that “while terminological investigations can certainly be carried out in a monolingual setting, one of its most widely practised applications is in the domain of translation.”

Terminology management systems are often integrated in computer-aided translation tools so that the contents of a term bank can be searched automatically during translation. In recent translation memory surveys (Lommel 2004, Lagoudaki 2006), respondents mention term replacements and terminology management as some of the reasons for using technology.

Due to the rapidly evolving technology and the language describing those quickly evolving specialized technological fields, terminology management is becoming increasingly important. Although free-to-use multilingual term banks (e.g. IATE[1], TermiumPlus[2], EuroTermBank[3], and TermSciences[4]) are available to translators, they usually include terms from a wide range of specialized fields and cannot keep pace with the rapidly evolving technological fields. Moreover, clients may prefer different terms to the ones included in the publicly available term banks. Nor do all of the above-mentioned term banks support smaller languages. For that reason, terminological work will always be part of the translator's workload.

Terms consist of single-words or multi-word units that represent discrete conceptual entities, properties, activities or relations in a particular domain (Bowker 2008). Two different theoretical viewpoints regarding the acquisition of terminology exist. While the early approaches to terminology were onomasiological (starting from the concepts and working towards the terms), the more recent corpus-driven approaches are per definition semasiological (starting from the terms and working towards the concepts). The different theories of terminology are described in (Cabré Castellví 2003) and (Bowker 2008).

Wright (1997) views the process of terminology management as an iterative one:

> Generally speaking, individual terminologists and even standardization committees begin their work in a descriptive fashion, collecting terms from existing texts and spoken discourse. [...] The terms are then used to create initial concept systems, and definitions are revised and redrafted based on relations revealed in the concept systems. (Wright 1997, p. 22)

Terminology extraction can be seen as an important step of a larger process: corpus compilation, terminology extraction and terminology management (Gamper and Stock 1999). In the terminology extraction phase, terms are identified in a text and – in the case of multilingual terminology extraction – the corresponding translations are retrieved. The extracted terms and their translations can be stored in bilingual glossaries, which are already a valuable aid for technical translators. If the aim is the creation of a term bank, the extracted terms are structured in concept-oriented databases in the terminology management phase. Each database entry represents a concept and contains all extracted term variants (including synonyms and acronyms) in several languages.

[1] http://iate.europa.eu

[2] http://www.btb.termiumplus.gc.ca

[3] http://www.eurotermbank.com

[4] http://www.termsciences.fr

It goes without saying that compiling terminology is a labour-intensive process. Therefore, computer-assisted terminology acquisition will definitely lead to increased efficiency. In most cases, the terminology extraction tool generates lists of candidate terms, which are then verified by human experts.

7.2 Terminology extraction approaches

Basically, there are two methodologically different approaches to terminology extraction, viz. the *linguistic* and the *statistical* approach. The linguistic approach is based on the characteristics of term formation patterns, which are expressed as part-of-speech code sequences (e.g. N N, N prep N, Adj N). Term formation patterns for English can be found in (Justeson and Katz 1995) and (Quin 1997); patterns for French in (Daille 1996).

In English, French, Dutch and Italian, – the languages we worked with – the majority of multi-word terms are formed by concatenating nouns, adjectives, prepositions and adverbs. However, different trends can be observed: in English, the most frequently used compounding strategy is the combination of nouns, while in French and Italian prepositional phrases are more often used; in Dutch, compounds are often not separated by means of white space characters and hence constitute single-word terms. In order to determine the morpho-syntactic patterns, linguistically-based systems apply language-specific part-of-speech taggers. As a result, linguistically-based terminology extraction programs are always language-dependent.

The statistical approach on the other hand is language-independent and is based on quantifiable characteristics of term usage. One such characteristic is that terms tend to occur more frequently in specialized texts than in general domain texts. The statistical approaches use several statistical measures such as frequency, association scores, diversity and distance metrics (Daille 1996).

Fulford (2001, p. 261) pointed out that “terms do not tend to possess linguistic features that distinguish them clearly and decisively from non-terms”. Hence, we can expect that linguistically-based approaches tend to overgenerate. Also the statistical approaches tend to produce some noise. Moreover, frequency-based systems will not be able to detect newly-coined terms with low frequencies.

Hence, most state-of-the-art systems use hybrid approaches that combine linguistic and statistical information. Different methods and systems are described and compared in Kageura and Umino (1996), Cabré Castellví, Bagot and Palatresi (2001), and Zhang, Iria, Brewster and Ciravegna (2008).

Bilingual term extraction is faced with the additional problem of finding translation equivalents in parallel texts. There is a long tradition of research into bilingual terminology extraction (Kupiec 1993, Gaussier 1998). In most systems, candidate terms are first identified monolingually. In a second step, the translation candidates are extracted from the bilingual corpus on the basis of word alignments or co-occurrence information. In recent work, Itagaki et al. (2007) use the phrase table derived from the GIZA++ alignments to identify the translations.

We present an alternative and more flexible approach that generates candidate terms directly from the aligned anchor chunks (see chapter 5). Rather than predefining terms as sequences of PoS patterns, we first generate candidate terms starting from the aligned phrases. In a second step, we use frequency information of a general purpose corpus and the n-gram frequency of the specialized text corpus to determine the term specificity.

Our approach is related to that of Tiedemann (2001) who uses the bilingual word alignment to improve the precision of the terms extracted in the monolingual term extraction phase. The main difference between his work and ours, however, is that Tiedemann starts from monolingual term extraction, while we take as a starting point the aligned linguistically motivated phrases.

7.3 Background

The bilingual term extraction module described here has been carried out in the framework of a terminology management project for PSA Peugeot Citroën, a major automotive company. The final goal of the project was a reduction and terminological unification process of the company's database, which contains all text strings that are used for compiling user manuals. French being the source language, all French entries had been translated to some extent into the twenty different languages that are part of the customer's portfolio. Bilingual term extraction was the first step of the more extensive terminology management project.

This chapter describes how we applied our novel terminology extraction module to the French-English, French-Italian and the French-Dutch part of the database[5].

[5] This chapter presents joint work with Els Lefever and Véronique Hoste and was published in Macken, Lefever and Hoste (2008) and in Lefever, Macken and Hoste (2009).

7.4 Corpus

The French database contains about 400,000 entries (i.e. sentences and parts of sentences with an average length of 9 words), which are aligned across all languages by means of a unique ID. For the development of the alignment and terminology extraction module, we created three parallel corpora: French-Italian, French-English, and French-Dutch. Details on the size of each parallel corpus can be found in table 7.1.

	Target Lang.	# Sentence pairs	# words
French	Italian	364,221	6,408,693
French	English	363,651	7,305,151
French	Dutch	364,311	7,100,585

Table 7.1: Number of sentence pairs and the total number of words in the three parallel corpora.

We PoS-tagged and lemmatized the French, English and Italian corpora with the freely available Treetagger tool (Schmid 1994). To annotate the Dutch corpus we used TADPOLE (Van den Bosch et al. 2007).

We mapped the part-of-speech tag sets of the different languages and created a language-independent version of the rule-based chunker described in chapter 2. The following example shows a French-English sentence pair divided in non-overlapping and non-recursive chunks:

> Fr: valable | uniquement | pour la ceinture | de sécurité avant latérale
> | du côté passager
>
> En: applies | only | to the outer seat belt | on the passenger side

Since we presume that sentence length has an impact on the alignment performance, and thus on term extraction, we created three test sets with varying sentence lengths. We distinguished short sentences (2-7 words), medium-length sentences (8-19 words) and long sentences (> 19 words). Each test corpus contains approximately 9,000 words; the number of sentence pairs per test set can be found in table 7.2. We also created a training corpus of approximately 5,000 words with sentences of varying length to debug the linguistic processing and the alignment module as well as to define the thresholds for the statistical filtering of the candidate terms (see section 7.6).

	# Words	# Sentence pairs
Short (< 8 words)	± 9,000	823
Medium (8-19 words)	± 9,000	386
Long (> 19 words)	± 9,000	180
Training corpus	± 5,000	393

Table 7.2: Number of words and sentence pairs in the test and training corpora.

7.5 Sub-sentential alignment

As the basis of our terminology extraction system, we used the anchor chunk alignment module described in chapter 5 that links linguistically motivated phrases in parallel texts based on lexical correspondences and syntactic similarity. We assume that these anchor chunks offer a valid and language-independent alternative to identify candidate terms on the basis of predefined PoS patterns. As the corpus of PSA Peugeot Citroën contains rather literal translations, we expect that a high percentage of anchor chunks will be retrieved.

Although the architecture of the sub-sentential alignment system is language-independent, some language-specific resources are used. First, a bilingual lexicon to generate the lexical correspondences and second, tools to generate additional linguistic information (a PoS tagger, a lemmatizer and a chunker). The sub-sentential alignment system takes as its input sentence-aligned texts, together with the additional linguistic annotations for the source and the target texts.

The source and target sentences are divided into chunks on the basis of PoS information, and lexical correspondences are retrieved from a bilingual dictionary. In order to extract bilingual dictionaries from the three parallel corpora, we used the Perl implementation of IBM Model One that is part of the Microsoft Bilingual Sentence Aligner (Moore 2002).

We followed the procedure described in chapter 5 to generate the bilingual dictionaries. For each parallel corpus, we ran IBM Model One in two directions: from source to target and from target to source. To get high-accuracy links, only the words pairs occurring in both the source-to-target and target-to-source word lists were retained, and the probabilities were averaged. To get rid of the noise produced by the translation model, only the entries with an averaged value of at least 0.1 were retained. This value was set experimentally[6].

[6] Lowering this threshold significantly decreased precision scores of the sub-sentential alignment system.

The resulting bilingual dictionaries contain 25,035 French-English word pairs, 30,532 French-Dutch word pairs and 27,497 French-Italian word pairs. The bilingual dictionaries were used to create the lexical link matrix for each sentence pair.

In order to link chunks on the basis of lexical clues and chunk similarity, the following steps are taken for each sentence pair:

1. Creation of the lexical link matrix.
2. Linking chunks on the basis of lexical correspondences and chunk similarity.
3. Linking remaining chunks.

For each source and target word, all translations for the word form and the lemma are retrieved from the bilingual dictionary. In the process of building the lexical link matrix, function words are neglected. For all content words, a lexical link is created if a source word occurs in the set of possible translations of a target word, or if a target word occurs in the set of possible translations of the source word. Identical strings in the source and target sentence are also linked.

The method to link anchor chunks is largely similar to that described in chapter 5. Candidate anchor chunks are selected on the basis of the information available in the lexical link matrix. The candidate target chunk is built by concatenating all target chunks from a *begin index* until an *end index*. The begin index points to the first target chunk with a lexical link to the source chunk under consideration. The end index points to the last target chunk with a lexical link to the source chunk under consideration. In this way, 1:1 and 1:n candidate target chunks are built. The process of selecting candidate chunks as described above, is performed a second time starting from the target sentence. In this way, additional n:1 candidates are constructed. For each selected candidate pair, a *similarity test* is performed. Chunks are considered to be similar if at least a certain percentage of words from the source and target chunk(s) are either linked by means of a lexical link or can be linked on the basis of corresponding part-of-speech codes. All word classes can be linked on the basis of corresponding PoS codes.

In addition to linking words on the basis of PoS codes, a small set of predefined language-dependent rules was implemented to handle function words:

- Additional function words (determiners and prepositions) in the source or target language are linked together with their noun and then refer to the noun's translation.
- In French, the preposition *de* is contracted with the definite articles *le* and *les* to *du* and *des* respectively. The contracted determiners are linked to an English or Dutch preposition and determiner. A similar phenomenon occurs in Italian.

The percentage of words that had to be linked was empirically set at 85%.

In a second step, chunks consisting of one function word – mostly punctuation marks and conjunctions – are linked on the basis of corresponding part-of-speech codes if their left or right neighbour on the diagonal is an anchor chunk. Corresponding final punctuation marks are also linked.

In a final step, additional candidates are constructed by selecting non-anchor chunks in the source and target sentence that have corresponding left and right anchor chunks as neighbours. The anchor chunks of the first step are used as contextual information to link n:m chunks or chunks for which no lexical link was found in the lexical link matrix.

In figure 7.1, the chunks [Fr: *gradient*] ∼ [En: *gradient*] and the final punctuation mark have been retrieved in the first step as anchor chunks. In the last step, the n:m chunk [Fr: *de remontée pédale d' embrayage*] ∼ [En: *of rising of the clutch pedal*] is selected as candidate anchor chunk because it is enclosed within anchor chunks.

Since the contextual clues (the left and right neighbours of the additional candidate chunks are anchor chunks) provide some extra indication that the chunks can be linked, the similarity test for the final candidates was somewhat relaxed: the percentage of words that had to be linked was lowered to 0.80 and a more relaxed PoS matching function was used.

To test the sub-sentential alignment module, we manually indicated all translational correspondences in the three test corpora.

We adapted the annotation guidelines described in chapter 3 to the French-English, French-Dutch and French-Italian language pairs, and used three different types of links: *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. Figure 7.2 shows an example for French-English.

We used Alignment Error Rate (see section 4.3) as evaluation metric to assess the system's performance. Table 7.3 shows the alignment results for the three language pairs.

	Short			Medium			Long		
	Prec	Rec	AER	Prec	Rec	AER	Prec	Rec	AER
Italian	.99	.93	.04	.95	.89	.08	.95	.89	.07
English	.97	.91	.06	.95	.85	.10	.92	.85	.12
Dutch	.96	.83	.11	.87	.73	.20	.87	.67	.24

Table 7.3: Precision, recall, and alignment error rate for our sub-sentential alignment system evaluated on the French-Italian, French-English and French-Dutch test corpora.

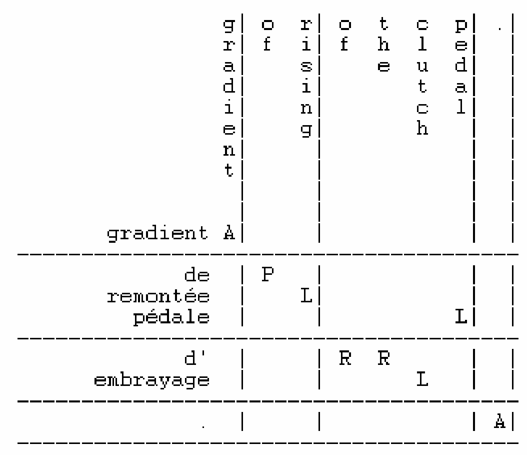
As expected, the results show that the alignment quality is closely related to the similarity between languages. As shown in example 17, Italian and French are syntactically almost identical – and hence easier to align, English and French are still close but show some differences (e.g. a different compounding strategy and word order) and French and Dutch present a very different language structure (e.g. in Dutch the different compound parts are not separated by spaces, *separable verbs* – verbs with prefixes that are stripped off – occur frequently (e.g. *losmaken* as an infinitive versus *maak los* in the conjugated forms) and a different word order is adopted).

(17) Fr: déclipper le renvoi de ceinture de sécurité [En literal translation: *unclip the mounting of the belt of safety*]

It: sganciare il dispositivo di riavvolgimento della cintura di sicurezza [En literal translation: *unclip the mounting of the belt of safety*]

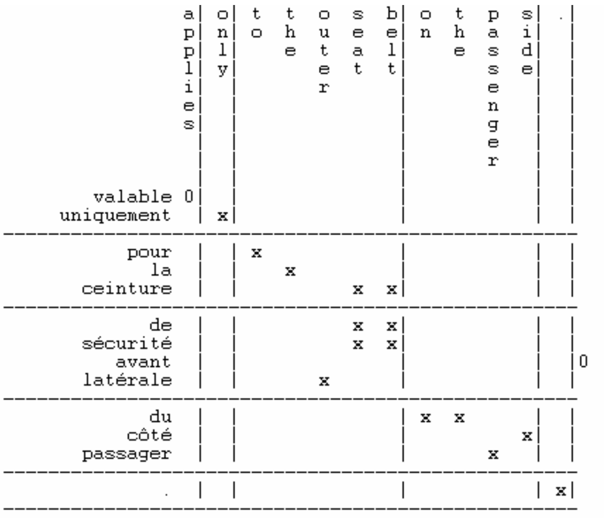
Nl: maak de oprolautomaat van de autogordel los [En literal translation: *clip the mounting of the seat-belt un*]
En: unclip the seat belt mounting

We tried to improve the low recall for French-Dutch by adding a decompounding module to our alignment system. If the Dutch word does not have a lexical correspondence in the French sentence, we decompose the Dutch word into its meaningful parts and look for translations of the compound parts. This implies that, without decompounding, in example 18 only the correspondences *doublure*



```
                g| o  r| o  t  c  p| .|
                r| f  i| f  h  l  e|  |
                a|    s|    e  u  d|  |
                d|    i|       t  a|  |
                i|    n|       c  l|  |
                e|    g|       h   |  |
                n|     |           |  |
                t|     |           |  |
                 |     |           |  |
                 |     |           |  |
       gradient A|     |           |  |
---------------------------------------
             de  | P   |           |  |
       remontée  |    L|           |  |
         pédale  |     |          L|  |
---------------------------------------
             d'  |     | R  R      |  |
      embrayage  |     |       L   |  |
---------------------------------------
              .  |     |           | A|
---------------------------------------
```

Figure 7.1: N:m candidate chunk: 'A' stands for anchor chunks, 'L' for lexical links, 'P' for words linked on the basis of corresponding PoS codes and 'R' for words linked by language-dependent rules.



```
                a| o| t  t  o  s  b| o  t  p  s| .|
                p| n| o  h  u  e  e| n  h  a  i|  |
                p| l|    e  t  a  l|    e  s  d|  |
                l| y|       e  t  t|       s  e|  |
                i|  |       r      |       e   |  |
                e|  |              |       n   |  |
                s|  |              |       g   |  |
                 |  |              |       e   |  |
                 |  |              |       r   |  |
                 |  |              |           |  |
        valable 0|  |              |           |  |
     uniquement  | x|              |           |  |
---------------------------------------------------
           pour  |  | x            |           |  |
             la  |  |    x         |           |  |
       ceinture  |  |          x  x|           |  |
---------------------------------------------------
             de  |  |          x  x|           |  |
       sécurité  |  |          x  x|           |  |
          avant  |  |              |           |  |0
       latérale  |  |       x      |           |  |
---------------------------------------------------
             du  |  |              | x  x      |  |
           côté  |  |              |          x|  |
       passager  |  |              |       x   |  |
---------------------------------------------------
              .  |  |              |           | x|
---------------------------------------------------
```

Figure 7.2: Manual reference: regular links are indicated by x's, fuzzy links and null links by 0's.

∼ *binnenpaneel*, *arc* ∼ *dakversteviging* and *arrière* ∼ *achter* will be found. By decomposing the compound into its meaningful parts (*binnenpaneel* = *binnen* + *paneel*, *dakversteviging* = *dak* + *versteviging*) and retrieving the lexical links for the compound parts, we were able to link the missing correspondence: *pavillon* ∼ *dakversteviging*.

(18) Fr: doublure arc pavillon arrière [En: rear roof arch lining]
Nl: binnenpaneel dakversteviging achter

We experimented with the decompounding module of Vandeghinste (2008, p. 71–82), which is based on the Celex lexical database (Baayen et al. 1993). The module, however, did not adapt well to the highly technical automotive domain, which was reflected by its low recall and the low confidence values for many technical terms. In order to adapt the module to the automotive domain, we implemented a domain-dependent extension to the decompounding module on the basis of the training corpus. This was done by first running the decompounding module on the Dutch sentences to construct a list with possible *compound heads*, being valid compound parts in Dutch. This list was updated by inspecting the decompounding results on the training corpus. While decompounding, we go from right to left and strip off the longest valid part that occurs in our preconstructed list with compound parts and we repeat this process on the remaining part of the word until we reach the beginning of the word.

Table 7.4 shows the impact of the decompounding module, which is more prominent for short and medium-length sentences than for long sentences. Decompounding – as described above – was implemented as a fall-back solution. When no lexical correspondences could be retrieved from the Model One dictionary, a second dictionary look-up was performed for the compound parts. However, as the Model One dictionary was derived from a corpus of word forms, the dictionary only contained Dutch compound parts that were also present in the corpus as separate words. We expect that splitting the compounds in the Dutch part of the parallel corpus prior to statistical word alignment might have a larger impact on the results.

	Short			Medium			Long		
	Prec	Rec	AER	Prec	Rec	AER	Prec	Rec	AER
Dutch									
no decomp	.95	.76	.16	.88	.67	.24	.88	.64	.26
decomp	.96	.83	.11	.87	.73	.20	.87	.67	.24

Table 7.4: Precision, recall, and Alignment Error Rate for French-Dutch without and with decompounding information.

7.6 Terminology extraction

As described at the end of section 7.2, we generated candidate terms from the aligned phrases. In a second step, we use a general-purpose corpus and the n-gram frequency of the automotive corpus to determine the specificity of the candidate terms.

We start from the anchor chunks as they are the minimal chunks that could be linked together. Two heuristics are used to generate candidate terms from the aligned anchor chunks: a first heuristic strips off adjectives and a second one considers each N + PP pattern as candidate term.

The following candidate terms are generated for example (19):

(19) Fr: une procédure d' initialisation du calculateur de boîte de vitesses automatique
En: an automatic gearbox ECU initialisation procedure

1	procédure d' initialisation du calculateur de boîte de vitesses automatique	automatic gearbox ECU initialisation procedure
2	procédure d' initialisation du calculateur de boîte de vitesses	gearbox ECU initialisation procedure
3	procédure d' initialisation	initialisation procedure
4	initialisation du calculateur	ECU initialisation
5	calculateur de boîte de vitesses	gearbox ECU
6	boîte de vitesses automatique	automatic gearbox
7	boîte de vitesses	gearbox
8	procédure	procedure
9	initialisation	initialisation
10	calculateur	ECU
11	automatique	automatic

As our terminology extraction module is meant to generate a bilingual automotive lexicon, every entry in our lexicon should refer to a concept or action that is relevant in an automotive context. This also means we want to include the minimal semantical units (e.g. seat belt) as well as the larger semantical units (e.g. outer front seat belt) of a parent-child term relationship.

To filter our candidate terms, we keep the following criteria in mind:

- Each entry in the extracted lexicon should refer to an object or action that is relevant for the domain. This criterion reflects the notion of *termhood* that is used to express "the degree to which a linguistic unit is related to domain-specific context" (Kageura and Umino 1996, pp. 260–261).

- Multiword terms should present a high degree of cohesiveness. This criterion reflects the notion of *unithood* that expresses the"degree of strength or stability of syntagmatic combinations or collocations" (Kageura and Umino 1996, pp. 260–261).

- All term pairs should contain valid translation pairs. So translation quality is also taken into consideration.

Two different statistical measures were used to determine *termhood* and *unithood*. To measure the termhood criterion and to filter out general vocabulary words, we applied Log-Likelihood filters on the French single-word terms. To measure unithood, we calculated Mutual Expectation for the multi-word terms in both the source and target language.

The Log-Likehood ratio should allow us to detect single-word terms that are *distinctive* enough to be kept in our bilingual lexicon (Daille 1996). This metric considers word frequencies weighted over two different corpora (in our case a technical automotive corpus and the more general purpose corpus *Le Monde*[7]), and assigns high Log-Likelihood values to words having much higher or lower frequencies than expected.

In the formula below, N corresponds to the number of words in the corpus, whereas the *observed values* O correspond to the real frequencies of a word in the corpus. The formula for calculating both the expected values (E) and the Log-Likelihood have been described in detail by Rayson and Garside (2000).

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{7.1}$$

We used the resulting Expected values for calculating the Log-Likelihood:

$$-2 log \lambda = 2 \sum_i O_i log(\frac{O_i}{E_i}) \tag{7.2}$$

Manual inspection of the Log-Likelihood figures confirmed our hypothesis that more domain-specific terms in our corpus were assigned high Log-Likelihood values. We experimentally defined the threshold for Log-Likelihood values corresponding to distinctive terms on our training corpus. Example (20) shows some translation pairs that are filtered out by applying the Log-Likelihood threshold.

[7] http://catalog.elra.info/product_info.php?products_id=438

(20) Fr: cependant – En: however – It: tuttavia – Nl: echter
Fr: choix – En: choice – It: scelta – Nl: keuze
Fr: continuer – En: continue – It: continuare – Nl: verdergaan
Fr: cadre – En: frame – It: cornice – Nl: frame (erroneous filtering)
Fr: allégement – En: lightening – It: alleggerire – Nl: verlichten (erroneous filtering)

The Mutual Expectation measure as described by Dias and Kaalep (2003) is used to measure the degree of cohesiveness between words in a text. In this way, candidate multi-word terms whose components do not occur together more often than expected by chance get filtered out. In a first step, we calculated all n-gram frequencies (up to 8-grams) for our four automotive corpora and then used these frequencies to derive the Normalised Expectation (NE) values for all multi-word entries, as specified by the formula of Dias and Kaalep:

$$NE = \frac{prob(n\text{-}gram)}{\frac{1}{n}\sum prob(n\text{-}1\text{-}grams)} \tag{7.3}$$

The Normalised Expectation value expresses the cost – in terms of cohesiveness – of the possible loss of one word in an n-gram. The higher the frequency of the n-1-grams, the smaller the NE, and the smaller the chance that it is a valid multi-word expression. The final Mutual Expectation (ME) value is then obtained by multiplying the NE values by the n-gram frequency. This means that the Mutual Expectation between n words in a multi-word expression is based on the Normalised Expectation and the relative frequency of the n-gram in the corpus.

We calculated Mutual Expectation values for all candidate multi-word term pairs and filtered out incomplete or erroneous terms having ME values below an experimentally set threshold (being below 0.005 for both source and target multi-word or below 0.0002 for one of the two multi-words in the translation pair). An example of incomplete candidate terms that were filtered out by applying the ME filter is presented in (21) :

(21) Fr: fermeture embout - En: end closing - It: chiusura terminale - Nl: afsluiting deel
(should be: Fr: fermeture embout de brancard - En: chassis member end closing panel - It: chiusura terminale del longherone - Nl: afsluiting voorste deel van langsbalk)

7.7 Evaluation

To evaluate the terminology extraction module, we used all the sentences of the three test corpora (see table 7.2). The output of the terminology extraction module was manually labelled by two PSA project staff members of the Faculty of Translation Studies of University College Ghent who were familiar with the terminology of PSA Peugeot Citroën. The annotators were asked to judge both the *translational quality* of the entry (both languages should refer to the same referential unit) as well as the *relevance* of the term in an automotive context. Three labels were used: OK (valid entry), NOK (not a valid entry) and MAYBE (when the annotator was not sure about the relevance of the term).

First, the impact of the statistical filtering was measured on the bilingual term extraction. Secondly, we compared the output of our system with the output of a commercial bilingual terminology extraction module and with the output of a set of standard monolingual term extraction modules.

Since the annotators labelled system output, only precision is calculated. In future work, we plan to develop a Gold Standard corpus, which will enable us to calculate recall.

7.7.1 Impact of filtering

	Not Filtered			Filtered		
	ok	Nok	Maybe	ok	Nok	Maybe
FR-EN						
Sing_w	82.0%	17.0%	1.0%	86.5%	12.0%	1.5%
Mult_w	81.0%	16.5%	2.5%	83.0%	14.5%	2.5%
FR-IT						
Sing_w	80.5%	19.0%	0.5%	84.5%	15.0%	0.5%
Mult_w	69.0%	30.0%	1.0%	72.0%	27.0%	1.0%
FR-NL						
Sing_w	72.0%	25.0%	3.0%	75.0%	22.0%	3.0%
Mult_w	83.0%	15.0%	2.0%	84.0%	14.0%	2.0%

Table 7.5: Impact of statistical filters on single and multi-word terminology extraction for three different language pairs.

Table 7.5 shows the difference in performance for both single and multi-word terms with and without filtering. Single-word filtering seems to have a higher impact on the results than multi-word filtering. This can be explained by the fact that our candidate multi-word terms are generated from anchor chunks (chunks aligned with a very high precision) that already adhere to strict syntactic constraints. The annotators also noted that especially for the single-word terms it was difficult to distinguish between technical and common vocabulary.

7.7.2 Comparison with bilingual terminology extraction

We compared the three resulting term lists (French-English, French-Dutch and French-Italian) with the output of a commercial state-of-the-art terminology extraction program SDL MultiTerm Extract[8]. MultiTerm Extract is a statistical terminology extraction system that first generates a list of candidate terms in the source language (French in our case) and then looks for translations of these terms in the target language. In order to allow for a fair comparison, we ran MultiTerm Extract with its default settings (default noise-silence threshold, default stopword list, and so on) on a large portion of our parallel corpus that also contains all test sentences[9].

Table 7.6 shows that even after applying statistical filters, our term extraction module retains a much higher number of candidate terms than MultiTerm Extract.

	# Before filtering	# After filtering	MultiTerm
FR-EN	4,052	3,386	1,831
FR-IT	4,381	3,601	1,704
FR-DU	3,285	2,662	1,637

Table 7.6: Number of extracted terms before and after applying Log-Likelihood and Mutual Expectation filters for the three language pairs and the number of extracted terms by MultiTerm Extract.

Table 7.7 presents the results of both systems and shows the differences in performance for single and multi-word terms.

[8] www.translationzone.com/en/products/sdlmultitermextract

[9] 70,000 sentences seemed to be the maximum size of the corpus that could be easily processed within MultiTerm Extract.

	Anchor chunk approach			Multiterm		
	ok	Nok	Maybe	ok	Nok	Maybe
FR-EN						
Sing_w	86.5%	12.0%	1.5%	77.0%	21.0%	2.0%
Mult_w	83.0%	14.5%	2.5%	47.0%	51.0%	2.0%
Total	84.5%	13.5%	2.0%	64.0%	34.0%	2.0%
FR-IT						
Sing_w	84.5%	15.0%	0.5%	85.0%	14.0%	1.0%
Mult_w	72.0%	27.0%	1.0%	65.0%	34.0%	1.0%
Total	77.5%	22.0%	1.0%	76.5%	22.5%	1.0%
FR-DU						
Sing_w	75.0%	22.0%	3.0%	64.5%	33.0%	2.5%
Mult_w	84.0%	14.0%	2.0%	49.5%	49.5%	1.0%
Total	79.5%	20.0%	2.5%	58.0%	40.0%	2.0%

Table 7.7: Precision figures for our term extraction system and for MultiTerm Extract

The following observations can be made:

- The performance of both systems is comparable for the extraction of single-word terms, but our system clearly outperforms MultiTerm Extract when it comes to the extraction of more complex multi-word terms.

- Although the alignment results for French-Italian were very good, we did not achieve comparable results for Italian multi-word extraction. This may be due to the fact that both languages are very similar at the syntactic level. As a result, smaller syntactic chunks are linked. However one can argue that, just because of the syntactic resemblance in both languages, the need for complex multi-word terms is less prominent in closely related languages as translators can just paste smaller noun phrases together in the same order in both languages. In example (22), we can recompose the larger compound *l'embout de brancard* or *il terminale del sottoporta* by translating the smaller parts in the same order *l'embout* ~ *il terminale* and *de brancard* ~ *del sottoporta*.

  (22) Fr: déposer | l' embout | de brancard
  It: togliere | il terminale | del sottoporta

- Despite the worse alignment results for Dutch, we achieve good accuracy results on the multi-word term extraction. Part of that can be explained by the fact that French and Dutch use a different compounding strategy: whereas French compounds are created by combining prepositional phrases, Dutch usually tends to concatenate nouns (even without inserting spaces between the different compound parts). As a results, we can extract larger Dutch chunks that correspond to several French chunks, for instance:

(23) Fr: feu régulateur | de pression carburant.
Nl: brandstofdrukregelaar.

7.7.3 Comparison with monolingual terminology extraction

In order to gain insights in the performance of our terminology extraction module without considering the validity of the bilingual term pairs, we contrasted our extracted English terms with the output of different state-of-the-art monolingual terminology extraction systems for English (Zhang et al. 2008), which are publicly available[10]. As we want to include both single words and multiword terms in our technical automotive lexicon, we only considered automatic term extraction systems that extract both categories.

We compared our system against 5 other term extraction systems:

1. Baseline system (Simple Term Frequency).
2. Weirdness algorithm (Ahmad, Gillam and Tostevin 1999) which compares term frequencies of the extraction and reference corpora.
3. C-value (Frantzi and Ananiadou 1999) which uses term frequencies as well as unit-hood filters (to measure the collocation strength of units) .
4. Glossex (Park, Byrd and Boguraev 2002, Kozakov et al. 2004) which uses term frequency information from both the extraction and the reference corpora and compares term frequencies with frequencies of the multi-word components.
5. TermExtractor (Sclano and Velardi 2007) which is comparable to Glossex but introduces the notion of "domain consensus" which "simulates the consensus that a term must gain in a community before being considered a relevant domain term".

[10] http://www.dcs.shef.ac.uk/∼ziqizhang

In all of the above algorithms, the English part of the automotive corpus is PoS tagged and linguistic filters (selecting nouns and noun phrases) are applied to extract candidate terms. In a second step, stopwords are removed and the same set of extracted candidate terms (1,105 single words and 1,341 multi-words) is ranked differently by each algorithm. Our filtered list of unique English automotive terms contains 1,279 single words and 1,879 multi-words in total. About 10% of the terms do not overlap between the two term lists.

To compare the performance of the ranking algorithms, we followed the procedure of Zhang et al. (2008) and selected the top terms (300 single and multi-word terms) produced by the above-mentioned algorithms and compared these with the top terms (300 single and multi-word terms) that are ranked by descending Log-Likelihood (calculated on the British National Corpus) and Mutual Expectation values. All candidate terms have been manually labeled by linguists. Table 7.8 shows the results of this comparison.

	Single Word terms			Multiword terms		
	ok	Nok	Maybe	ok	Nok	Maybe
Baseline	80.0%	19.5%	0.5%	84.5%	14.5%	1.0%
Weirdness	95.5%	3.5%	1.0%	96.0%	2.5%	1.5%
C-value	80%	19.5%	0.5%	94.0%	5.0%	1.0%
Glossex	94.5%	4.5%	1.0%	85.5%	14.0%	0.5%
TermExtr.	85.0%	15.0%	0.0%	79.0%	20.0%	1.0%
AC approach	85.5%	14.5%	0.0%	90.0%	8.0%	2.0%

Table 7.8: Results for monolingual term extraction on the English part of the automotive corpus

Although our term extraction module was tailored towards bilingual term extraction, the results look competitive to monolingual state-of-the-art term extraction systems. If we compare these results with our bilingual term extraction results, we can observe that we gain more in performance for multi-words than for single words, which might mean that the filtering and ranking on the basis of Mutual Expectation works better than the Log-Likelihood ranking.

Another striking observation is that the weirdness algorithm scores very high, both for single-word and multi-word terms. As mentioned above, the weirdness algorithm compares the term frequencies of the extraction corpus with the frequencies of the general-domain reference corpus. As most single-word terms in the PSA corpus are very specific, the weirdness algorithm performs well on single-word terms for this type of corpus.

In the implementation of Zhang et al. (2008), the weirdness of a multi-word term is the average of the weirdness values of each part of the multi-word term. A possible explanation for the high scores of the weirdness algorithm for the multi-word terms is that most English multi-word terms are formed by combining nouns only (without prepositions that also occur in a general-domain corpus), and that the parts of the noun compounds in the PSA corpus are rather specific words.

An error analysis of the output of all algorithms leads to the following insights:

- All systems suffer from partial retrieval of complex multi-words (e.g. *management ecu* instead of *engine management ecu*, *chassis leg end piece closure* instead of *chassis leg end piece closure panel*).
- We manage to extract nice sets of multi-words that can be associated with a given concept, which could be helpful for automatic ontology population (e.g. *gearbox casing, gearbox casing earth, gearbox casing earth cable, gearbox control, gearbox control cables, gearbox cover, gearbox ecu, gearbox ecu initialisation procedure, gearbox fixing, gearbox lower fixings, gearbox oil, gearbox oil cooler protective plug*).
- Sometimes smaller compounds are not extracted because they belong to the same syntactic chunk (e.g. *passenger compartment assembly, passenger compartment safety, passenger compartment side panel* are extracted, but *passenger compartment* is not).

7.8 Summary

In this chapter, we described how the sub-sentential alignment module was used to guide bilingual terminology extraction. Comparisons with standard terminology extraction programs show an improvement of up to 20% for bilingual terminology extraction and competitive results (85% to 90% accuracy) for monolingual terminology extraction. In the near future we want to experiment with other filtering techniques, especially to measure the domain distinctiveness of terms. We also plan to create Gold Standards for bilingual terminology extraction, which will allow us to compute recall next to precision.

We also demonstrated the language independence of our sub-sentential alignment approach as the terminology extraction experiments were carried out on three different language pairs, viz. French-English, French-Italian and French-Dutch. To that end, the part-of-speech code sets of the different languages were mapped and a language-independent version of the rule-based chunker described in chapter 2 was developed.

CHAPTER 8

Translation Memory Systems

8.1 Introduction

In the previous chapter, we demonstrated that the sub-sentential alignment system we developed is useful for bilingual terminology extraction. In this chapter, we will have a closer look at another application for human translators, viz. translation memory systems. More specifically, we will compare the performance of two commercially available systems: a sentence-based translation memory system (SDL Trados Translator's Workbench[1]) and a sub-sentential translation memory system (Similis[2]). We will evaluate the sub-sentential alignment module of Similis and compare its performance with that of our system[3].

Translation memory systems aim to reuse previously translated texts. The basic idea is fairly straightforward. Translation memory systems store source segments together with their translations in a database for reuse. During translation, the new text to be translated is segmented and each segment is compared with the source text segments of the database. When a useful match is found,

[1] http://www.trados.com

[2] http://www.lingua-et-machina.com

[3] The work described here is published in Macken (2009).

the retrieved source-target segment pair is provided to the translator. If no useful match is found, the translator translates the segment manually and the newly translated segment is added to the database.

Two processes in the above description are important for fully understanding the potential value and limitations of translation memory systems: segmentation and matching. In translation memory systems of the first generation[4], a segment corresponds to a sentence or a sentence-like unit such as a title, header or list item. The text is segmented on the basis of punctuation and document-formatting information. However, there is a major problem with the idea of using sentences as basic units of translation. Because the matching process is sentence-based, the potential value of the use of a translation memory system depends on the degree of full-sentence repetition of the text to be translated in the database. Consequently, translation memories are mainly used for translating technical documents (e.g. user manuals) or texts with related content (related products) or text revisions.

Several researchers have explored the idea of creating sub-sentential translation memories (Gotti et al. 2005, Planas and Furuse 2003, Simard and Langlais 2001). In this chapter, we compare the performance of a sentence-based translation memory system of the first generation with a sub-sentential translation memory system of the second generation. We then compare the translation suggestions made by the two different systems.

The second important process mentioned above is matching. During translation, the translation memory system matches the new source sentence with the source sentences in its database and proposes previously translated sentences to the translator. The system can either return sentence pairs with identical source segments (exact matches) or sentences that are similar but not identical to the sentence to be translated (fuzzy matches). In traditional translation memory systems, similarity is calculated by comparing surface strings, i.e. sequences of characters. In SDL Trados Translator's Workbench, the similarity threshold ranges from 30% to 99%. The user can change the similarity threshold in order to find the proper balance between precision and recall: if the similarity threshold is too high, potentially useful sentence pairs may be missed (high precision, low recall); if the similarity threshold is too low, the match can be based on high-frequency function words and the proposed translations may be of no use (low precision, high recall).

[4] The terms *first generation TM* and *second generation TM* are widely used to refer to sentence-based and sub-sentential translation memory systems respectively (Planas 2005, Lagoudaki 2008). Only Gotti et al. (2005) make another distinction: first generation systems are sentence-based translation memory systems without fuzzy matching techniques; second generation systems are sentence-based systems supporting fuzzy matches, and third generation systems are sub-sentential translation memory systems.

Because sentence-based translation memory systems calculate the similarity value on the whole surface string, sentence pairs that are very similar for humans may receive a low similarity value. Consider the following example sentences:

(24) Oracle® is a registered trademark of Oracle Corporation.

(25) Java® is a registered trademark of Sun Microsystems Inc.

(26) Unix®, X/Open®, OSF/1®, and Motif® are registered trademarks of the Open Group.

For a human it is obvious that example sentences 25 and 26 are very similar to example sentence 24.

However, the translation memory system assigns a fuzzy match of 61% to the second sentence and a fuzzy match of less than 30% to the third. As these examples demonstrate, translation memories contain smaller segments than sentences that can be useful for translators.

In the following sections, we describe several experiments that were carried out to assess the usefulness of different types of translation memory systems. As we are unaware of any comparative study on the degree of repetitiveness in different text types, an experiment was set up to quantify the recurrence level of complete sentences in different text types.

We also compare the performance of a sentence-based translation memory system with a sub-sentential translation memory system on different text types. As an example of a first generation system, we use SDL Trados Translator's Workbench, which is, according to the LISA Translation Memory Survey (Lommel 2004), the most widely used Translation Memory tool. As an example of a second generation system, we use Similis. According to Lagoudaki (2008), only two commercially sub-sentential translation memory systems are available: Similis and Masterin. Because Masterin only supports English, Swedish and Finnish, we opted for Similis as sub-sentential translation memory system.

8.2 Corpus

Three subcorpora with parallel texts belonging to three domains and three different text types were selected from the Dutch Parallel Corpus (see section 2.2.3). For each subcorpus, approximately 50,000 words of sentence-aligned parallel text was used to populate the translation memory, and approximately 2,000 words of source-text material was selected as text to be translated:

- The *medical* subcorpus contains European Public Assessments Reports (EPARs) originating from one pharmaceutical company. The texts are rather technical with a clear, repetitive structure. The texts were translated from English into Dutch.

- The *financial* subcorpus consists of a collection of newsletters from a bank that provide financial news for investors. The texts were originally written in Dutch and translated into English.

- The *journalistic* subcorpus contains articles originally published in *The Independent* and translated into Dutch for *De Morgen*.

We expect the highest degree of repetitiveness in the medical subcorpus; the lowest in the journalistic subcorpus. The manually corrected sentence alignments available in the Dutch Parallel Corpus reveal that a different translation strategy was adopted for the medical and financial documents than for the journalistic texts (see table 8.1). In the medical and financial texts, most of the correspondences at sentential level are 1:1 alignments (98% and 97%, respectively). In the journalistic texts, the 1:1 alignments only account for 70%; 1:2 and 2:1 alignments for 11%; and null alignments (sentences that were added or deleted) for 16%.

Domain	0:n	n:0	1:1	1:2	2:1	n:m	Total
Medical	1	0	1,478	12	13	0	1,504
Financial	3	7	1,425	11	15	2	1,463
Journalistic	122	83	881	135	12	19	1,252

Table 8.1: Number of different types of sentence alignments as extracted from the DPC.

The selected source texts also differ in average sentence length: the average sentence length of the source texts is 16.3 words for the medical texts, 14.7 words for the financial texts and 21.5 words for the journalistic texts. As long sentences tend to be translated by more than one sentence, the difference in average sentence length explains the high degree of 1:2 alignments in the last text type. As translation memory systems first segment the texts into sentence-like units and look for matching segments in their databases, the different sentence-alignment characteristics already indicate that some text types (i.e. journalistic texts) are less suited for translation with translation memories.

8.3 Sentence-based translation memory

In our first experiment, we used SDL Trados Translator's Workbench, a sentence-based translation memory system of the first generation. We created three translation memories (one for each subcorpus) and populated the translation memories with the sentence-aligned parallel texts. The obtained translation memories are a reduced version of the parallel corpora, as only unique sentence pairs without empty source or target segment (non-null alignments) are retained. Table 8.2 presents an overview of the size of the translation memories and the reduction rate.

Domain	Translation memory
Medical	908 (60%)
Financial	1294 (88%)
Journalistic	1047 (83%)

Table 8.2: Size of the resulting translation memory actually used by SDL Trados Translator's Workbench.

A size reduction is seen in all three resulting translation memories, yet only for the medical and the financial translation memories is the reduction due to repetition at the sentence level. In the journalistic texts, the reduction is completely attributable to the removal of null alignments. We used the analysis function of SDL Trados Translator's Workbench to count the number of exact and fuzzy matches in the respective original source texts. During analysis, SDL Trados Translator's Workbench segments the source documents, compares the segments with the selected translation memory and examines the source document for text-internal repetition. The results are presented in tables 8.3, 8.4 and 8.5. Different match types are distinguished: text-internal repetitions (repetitions); exact matches (100%); and fuzzy matches within different threshold intervals (95-99%, 85-94%, 75-85% and 50-74%). For each match type, the second column contains the number of segments covered; the third column the total number of words; and the fourth column the percentage of the number of words covered.

The analysis statistics show that for 30% of the segments of the medical source texts, a translation suggestion is available in the translation memory. The percentage of translation suggestions drops to 9% for the financial texts, and not a single suggestion is available for the journalistic texts.

To assess the usefulness of the suggested translations, we pre-translated the source texts with a fuzzy match threshold at 70% and manually inspected the

Match Type	Number of segments	Number of words	Percentage
Repetitions	0	0	0
100%	17	236	13
95-99%	4	47	3
85-94%	11	126	7
75-84%	16	87	5
50-74%	2	35	2
No match	70	1,334	70
Total	120	1,865	100

Table 8.3: Analysis statistics (SDL Trados Translator's Workbench) for the medical texts.

Match Type	Number of segments	Number of words	Percentage
Repetitions	4	14	1
100%	10	74	3
95-99%	3	37	2
85-94%	1	12	1
75-84%	3	15	1
50-74%	1	27	1
No match	122	1,980	91
Total	144	2,159	100

Table 8.4: Analysis statistics (SDL Trados Translator's Workbench) for the financial texts.

translation suggestions. All suggested translations were considered to be either correct or useful, but the scope was considered limited:

- The EPARs (European Public Assessments Reports) of the medical subcorpus follow a clear, predefined structure. Apart from some introductory and closing paragraphs, the translation suggestions covered mainly the text headings, in which the name of a medicine was replaced (e.g. *What is the risk associated with <Xigris>?*).
- In the financial texts, the translation suggestions were only available for short headers and a few recurring paragraphs.
- In the journalistic texts, no translation suggestions were available.

From this small-scale experiment, we can conclude that some text types are

Match Type	Number of segments	Number of words	Percentage
Repetitions	1	1	0
100%	0	0	0
95-99%	0	0	0
85-94%	0	0	0
75-84%	0	0	0
50-74%	0	0	0
No match	126	1,981	100
Total	127	1,982	100

Table 8.5: Analysis statistics (SDL Trados Translator's Workbench) for the journalistic texts

more suited to be translated by means of a translation memory system than others. A second observation is that the analysis figures should be interpreted carefully. In the medical texts, the statistics indicate that 30% of the segments recur. However, manual inspection of the sentence-based translation suggestions showed that the impact was rather low.

8.4 Chunk-based translation memory

In our second experiment, we compared the performance of Similis, a commercially available sub-sentential translation memory system of the second generation, on the same test set. Similis is a linguistically enhanced translation memory in that it contains monolingual lexicons and chunkers to group words into phrases (Planas 2005). As a consequence, Similis is language-dependent. At present, Similis supports the following seven European languages: English, German, French, Italian, Spanish, Portuguese, and Dutch. Similis can be classified as a sub-sentential translation memory, as it can retrieve matches at the sub-sentential level. Translation memory systems working at the sub-sentential level face more challenges than sentence-based systems. In order to suggest matches at a sub-sentential level, the systems must be able to align source and target chunks (a non- trivial task); and must be able to identify (fuzzy) matches at sub-sentential level and have a mechanism to score multiple sub-sentential matches and select the best match.

In the following sections, we examine what type of structures Similis considers as chunks and we investigate the ability of Similis to align source and target chunks. In section 8.4.2, we evaluate the translation suggestions of Similis for our three text types; in section 8.4.3 we enlarge the size of the translation memories and examine how this affects our findings; in section 8.5 we compare the sub-sentential translation suggestions of Similis with the auto-concordance search of SDL Trados Translator's workbench.

8.4.1 Sub-sentential alignments in Similis

Similis does not only align sentences but also chunks below sentence level. In order to evaluate the quality of the aligned source and target chunks in Similis, we compared the aligned chunks with the manual reference described in section 3.6.

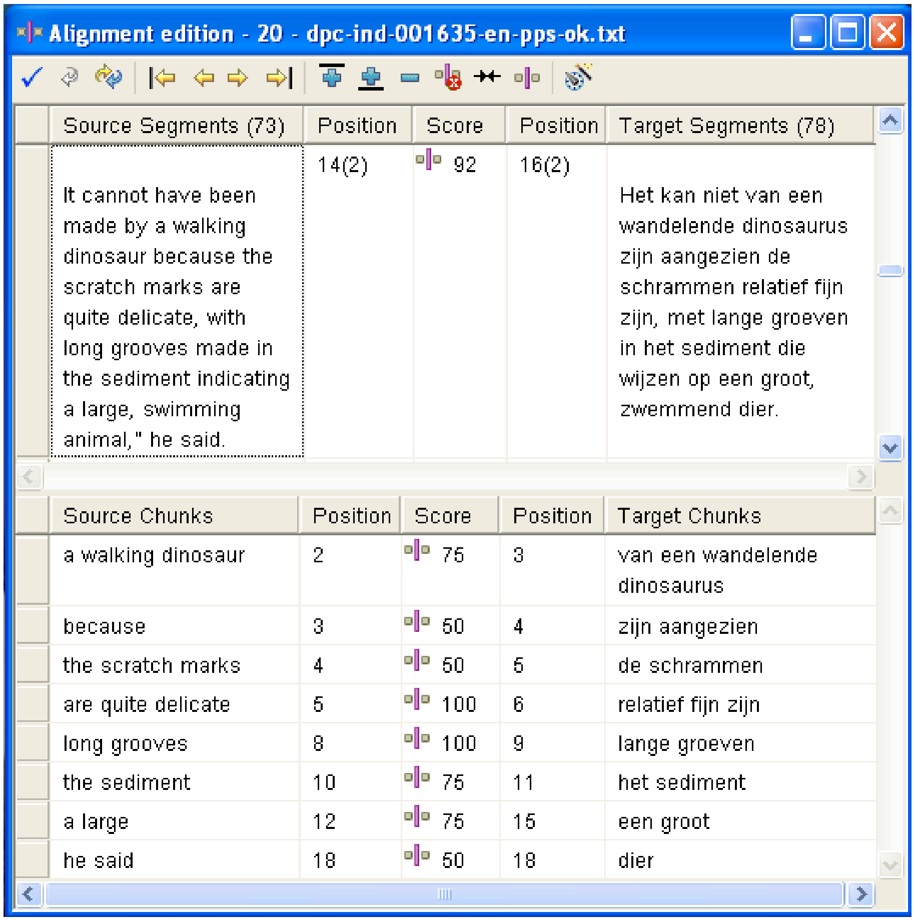
Similis defines a chunk as a phrase:

> SIMILIS does not only align sentences, but also chunks (or phrases) with their translations. A phrase is a structural unit in a text: a nominal or verbal group. It is defined by the grammatical categories of the words of which it is composed, and which are obtained by the linguistic analyser. A phrase is sometimes called *chunk*. (Similis, Guide de l'utlilisateur, version 2, p. 4)[5]

The *Edit Alignment* function of Similis allowed us to inspect the aligned chunks. As can be seen in figure 8.1, Similis's chunks can consist of sequences of several words, but one-word chunks also occur. Table 8.6 presents an overview of the number of source chunks of different lengths that were aligned by Similis in the three test corpora. The majority of aligned source chunks are relatively short chunks: over 50% consist of maximally two words, and 75% contain maximally three words.

Similis does not only store basic linguistic phrases, such as noun phrases (e.g. *the extinction of the dinosaurs* ∼ *het uitsterven van de dinosaurussen*), prepositional phrases (e.g. *into a vein* ∼ *in een ader*) and verb phrases (e.g. *were linked* ∼ *gelieerd zijn*), but also larger units in the translation memory (e.g. *the full list is available in the Package Leaflet* ∼ *zie de bijsluiter voor de volledige lijst van geneesmiddelen*). In most cases, these larger units are extracted from parenthetical expressions in the text.

[5] Original French citation: "SIMILIS met en correspondance non seulement les phrases mais aussi les chunks (ou syntagmes) avec leur traductions. Un syntagme est une unité structurelle du texte: un groupe nominal ou verbal. Il est défini grâce aux catégories grammaticales des mots qui le composent, et qui sont trouvées par l'analyseur linguistique. Un syntagme est parfois appelé *chunk*."



Source Segments (73)	Position	Score	Position	Target Segments (78)
It cannot have been made by a walking dinosaur because the scratch marks are quite delicate, with long grooves made in the sediment indicating a large, swimming animal," he said.	14(2)	92	16(2)	Het kan niet van een wandelende dinosaurus zijn aangezien de schrammen relatief fijn zijn, met lange groeven in het sediment die wijzen op een groot, zwemmend dier.

Source Chunks	Position	Score	Position	Target Chunks
a walking dinosaur	2	75	3	van een wandelende dinosaurus
because	3	50	4	zijn aangezien
the scratch marks	4	50	5	de schrammen
are quite delicate	5	100	6	relatief fijn zijn
long grooves	8	100	9	lange groeven
the sediment	10	75	11	het sediment
a large	12	75	15	een groot
he said	18	50	18	dier

Figure 8.1: Aligned source and target chunks for one sentence pair in Similis.

Source chunk size	Percentage
1	24
2	32
3	19
4	10
5	8
5-10	6
>10	0

Table 8.6: Size of the source chunks expressed in number of words and percentage in the test corpus.

We used the *Edit Alignment* function of Similis to collect all aligned source and target chunks and compared the aligned chunks with the manual reference. Each aligned chunk was given one of the following three labels:

- **Correct** if the aligned chunks were completely in line with the manually created reference alignment, e.g. *the scratch marks* ∼ *de schrammen* [En: *the scratch marks*].
- **Partially correct** if the source or target chunks contained extra words that were not aligned in the manually created reference alignment, e.g. *because* ∼ *zijn aangezien* [En: *been because*].
- **Wrong** if none of the words were aligned in the manually created reference alignment, e.g. *he said* ∼ *dier* [En: *animal*].

Table 8.7 summarizes the results of the analysis. The results demonstrate that word alignment (and hence chunk alignment) is a non-trivial task. For the medical texts, which are translated rather literally, 80% of the chunks are aligned correctly, and 3% are wrong alignments. For the financial texts, however, which are characterized by a high percentage of idiomatic expressions, and the journalistic texts, which are translated more freely, the percentage of correctly aligned chunks drops to 70% and 67%, respectively, and the percentage of wrongly aligned chunks rises to 5% and 7%, respectively. Applying fuzzy match techniques on an already error-prone translation memory can lead to quite unexpected results.

Domain	Correct	Partially	Wrong	Aligned words
Medical	80%	18%	2%	42%
Financial	70%	25%	5%	36%
Journalistic	67%	26%	7%	31%

Table 8.7: Percentage of correctly, partially correct or wrongly aligned chunks and percentage of aligned words.

8.4.2 Similis's translation suggestions

We used the analysis function of Similis to count the number of exact and fuzzy matches at segment and chunk level. The results are presented in table 8.8. The upper rows present segment matches, which roughly correspond to the statistics given by SDL Trados Translator's Workbench. Minor differences due to application of slightly different segmentation rules and a different calculation of the fuzzy-match scores can be observed.

Segment Match	Medical texts		Financial texts		Journalistic texts	
Match Type	Segments	Words	Segments	Words	Segments	Words
100%	12.6	12.3	14.5	5.7	3.2	0.2
95-99%	1.7	1.7	1.4	1.3	0.0	0.0
85-94%	18.5	8.5	1.4	0.6	0.0	0.0
75-84%	3.4	2.5	2.1	1.9	0.0	0.0
65-74%	2.5	2.2	0.0	0.0	0.0	0.0
<65%	0.0	0.0	0.0	0.0	0.0	0.0
Total	38.7	27.2	19.3	9.4	3.2	0.2
Chunk Match	Medical texts		Financial texts		Journalistic texts	
Match Type	Segments	Words	Segments	Words	Segments	Words
100%		2.1		4.3		0.5
95-99%		0.0		0.0		0.0
85-94%		9.5		7.3		3.6
75-84%		2.8		4.3		0.8
65-74%		1.8		1.7		1.3
<65%		0.0		0.0		0.0
Total		16.3		17.6		6.2

Table 8.8: Analysis statistics (Similis) for the three text types: percentage of segments and percentage of words per match type.

The lower rows present the additional matches at chunk level. As with the matches at segment level, matches at chunk level can be exact (100%) or fuzzy (ranging from 65–99%). Overall, the percentage of words for which sub-sentential translation suggestions are provided ranges from 16–17% (medical and financial texts) to 6% (journalistic texts).

Unfortunately, the statistics do not offer an indication as to the usefulness of the suggested translation. In many cases, the matched chunks are basic vocabulary words (e.g. *has* ∼ *heeft*, *that* ∼ *dat*, *came* ∼ *kwam*, *had* ∼ *had*, *more* ∼ *meer*, *now* ∼ *nu*, *worse* ∼ *erger*, *the world* ∼ *de wereld*) and are thus of no use to an experienced translator.

To assess the usefulness of the sub-sentential translation suggestions, we pre-translated the source texts, manually inspected all translation suggestions at sub-sentential level and assigned to each chunk one of the following three labels:

- **Basic vocabulary** if the matched chunk contained only basic vocabulary words.
- **Useful** if the matched chunk and translation suggestion contained some useful suggestion. The match could be a fuzzy match, and the proposed suggestion is not always entirely correct.
- **Wrong** if the proposed translation did not make sense due to alignment errors (see section 8.4.1).

The results are presented in table 8.9.

Domain	Basic Vocabulary	Useful	Wrong
Medical	15%	79%	6%
Financial	20%	78%	2%
Journalistic	54%	37%	9%

Table 8.9: Analysis of the sub-sentential translation suggestions.

We observe a high percentage of useful matches in the medical and financial texts and a low percentage of useful matches in the journalistic texts. This is because the medical and financial texts address similar topics and contain a high degree of recurring terms or recurring expressions. The journalistic articles have a more diverse content, and thus less recurring expressions.

However, the percentage of useful sub-sentential suggestions must be interpreted as an upper bound. There are two reasons for this. First, all sub-sentential translation suggestions were counted, not only the unique ones (e.g. in the financial texts, the word group *de aandelen* ∼ *the shares* occurred several times).

Second, whenever the translation suggestion made sense and did not belong to the basic vocabulary, the proposed translation was labelled as useful. However, the usefulness of most fuzzy matches at sub-sentential level is questionable. For example, for the word group *de Europese nutsbedrijven* [En: *the European utility companies*], a fuzzy match leads to a translation suggestion of *de Europese beurzen* [En: *the European stockmarkets*], which is hardly useful, as the translation difficulty is in the noun *nutsbedrijven* and not in the adjective *Europese*.

This limited experiment shows that the added value of the sub-sentential translation suggestions lays mainly in providing translation suggestions for terminology and frequent multiword expressions. Given the importance of terminology for the translation of domain-specific texts, the added value of using a sub-sentential translation memory system is considered to be high in such cases. Examples of useful suggestions from the financial domain are *portefeuille* [En: *portfolio*], *Duitse obligatierente* [En: *German bond rates*], *rentewapen* [En: *interest-rate weapon*], *bedrijfsinvesteringen* [En: *corporate investments*]. As demonstrated above, the usefulness of fuzzy matches on sub-sentential translation suggestions is less clear. A mechanism to filter out basic vocabulary words by for example using a high-frequency word list or using measures like TF-IDF (Sparck Jones 1979) would be beneficial.

8.4.3 Size of the translation memory

Because it is interesting to examine how the size of the translation memories affects our findings, we extracted additional parallel texts from the Dutch Parallel Corpus. We enlarged the translation memories from 50,000 words to 285,000 words of medical texts, 182,000 words of financial texts, and 289,000 words of journalistic texts. Table 8.10 presents the analysis results of Similis using the enlarged translation memories. The analysis statistics show that enlarging the translation memory has a positive effect at the level of segment matches for the financial texts: 18.6% exact matches versus 14.5% and 30.3% (all) matches versus 19.3%. Enlarging the translation memory has no effect at the level of segment matches for the jounalistic texts. For the medical texts, there is a slightly negative effect at the level of segment matches, but a positive effect at the level of chunk matches. It seems that if sentences contain fuzzy matches at both segment level and chunk level then the selection mechanism of Similis favours fuzzy matches with the highest threshold regardless of its type. For all text types, enlarging the translation memory has a positive effect on the chunk matches: 23.8% versus 16.3% for the medical texts; 22.2% versus 17.6% for the financial texts and 11.9% versus 6.2% for the journalistic texts.

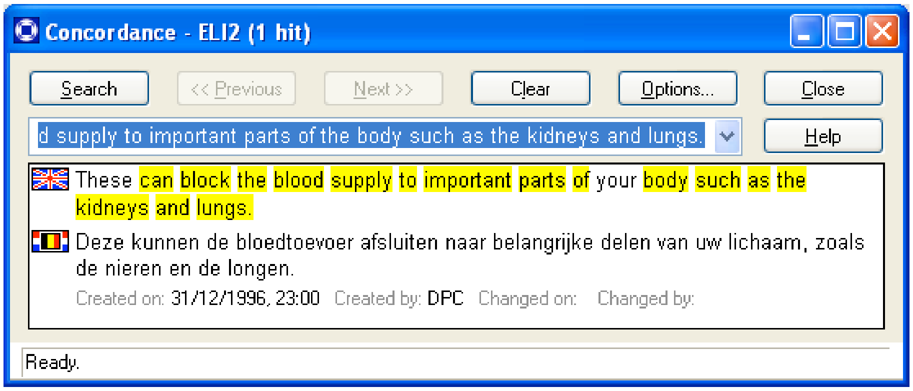
Segment Match	Medical texts		Financial texts		Journalistic texts	
Match Type	Segments	Words	Segments	Words	Segments	Words
100%	11.8	10.9	18.6	9.2	3.2	0.2
95-99%	1.7	1.7	2.8	3.1	0.0	0.0
85-94%	17.7	7.2	4.8	3.7	0.0	0.0
75-84%	3.4	2.5	2.1	1.9	0.0	0.0
65-74%	1.7	1.8	2.0	2.8	0.0	0.0
<65%	0.0	0.0	0.0	0.0	0.0	0.0
Total	36.1	24.1	30.3	20.6	3.2	0.2
Chunk Match	Medical texts		Financial texts		Journalistic texts	
Match Type	Segments	Words	Segments	Words	Segments	Words
100%		3.0		5.0		1.6
95-99%		0.0		0.0		0.1
85-94%		15.5		13.3		8.0
75-84%		3.5		2.7		1.9
65-74%		1.7		1.2		0.3
<65%		0.0		0.0		0.0
Total		23.8		22.2		11.9

Table 8.10: Analysis statistics (Similis) for the three text types using larger translation memories: percentage of segments and percentage of words per match type.

8.5 Autoconcordance (SDL Trados) versus subsentential translation suggestions (Similis)

SDL Trados Translator's Workbench also contains mechanisms to provide the translator with sub-sentential translation suggestions, viz. the auto-concordance search. If no match is found at segment level, the auto-concordance search retrieves all possible matches from the translation memory on the basis of the segment's lexical items and opens a concordance window showing all matching translation units. Figure 8.2 presents the auto-concordance result for the sentence "Excessive blood clotting is a problem during severe sepsis, when the blood clots can block the blood supply to important parts of the body such as the kidneys and lungs".

A drawback of the auto-concordance search is that the system searches for all matches, even when the translator may not need help with a particular passage. A second shortcoming is that the system does not align source and target chunks. The translator must scan the provided target sentence(s) to locate the correct translation suggestions. Moreover, the autoconcordance results are presented in



d supply to important parts of the body such as the kidneys and lungs.

These can block the blood supply to important parts of your body such as the kidneys and lungs.

Deze kunnen de bloedtoevoer afsluiten naar belangrijke delen van uw lichaam, zoals de nieren en de longen.

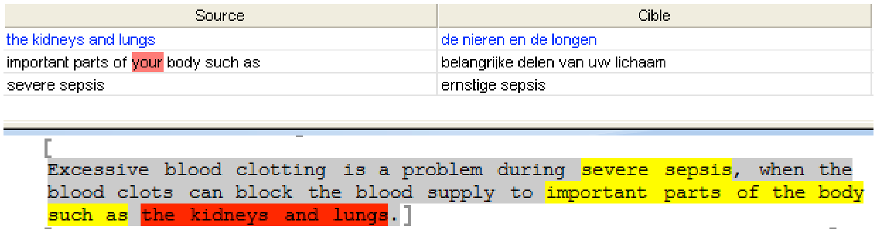
Created on: 31/12/1996, 23:00 Created by: DPC Changed on: Changed by:

Ready.

Figure 8.2: Autoconcordance result in SDL Trados Translator's workbench.

another window than the working window in which the translator is working.

Figure 8.3 shows how Similis presents sub-sentential matches to the user. In the sentence to be translated, the sub-sentential matches are indicated by colours. In the example, one exact match (*the kidneys and lungs* ∼ *de nieren en de longen*), and two fuzzy matches (*important parts of the body such as* ∼ *belangrijke delen van uw lichaam* and *severe sepsis* ∼ *ernstige sepsis*) are presented. Contrary to the auto-concordance function of SDL Trados Translator's Workbench, Similis presents the sub-sentential translation suggestions together with the segment matches in the translation environment. Visually, this is less distracting. Moreover, as Similis aligns source and target chunks, translation suggestions below sentence level are presented to the translator and she or he does not need to read an entire series of potentially useful target sentences.



Source	Cible
the kidneys and lungs	de nieren en de longen
important parts of your body such as	belangrijke delen van uw lichaam
severe sepsis	ernstige sepsis

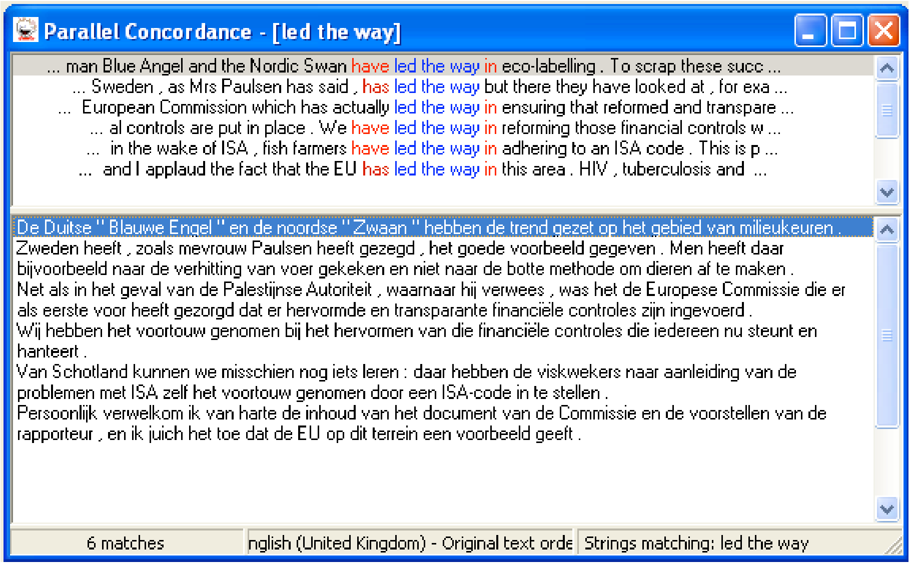
[Excessive blood clotting is a problem during severe sepsis, when the blood clots can block the blood supply to important parts of the body such as the kidneys and lungs.]

Figure 8.3: Sub-sentential translation suggestions provided by Similis.

8.6 Bilingual concordance tools

A remaining shortcoming of the current sub-sentential translation memory systems is that they fail to provide translation assistance for idiomatic expressions and collocations. Such expressions are not always contiguous and can appear in various forms in the texts. Because it is very difficult to align such expressions (idiomatic expressions are often not translated literally in the target language), sub-sentential translation suggestions are in most cases not available.

Luckily, for such expressions, a bilingual concordance tool such as Paraconc, which offers more powerful searches than the concordance function available in SDL Trados Translator's Workbench, may provide assistance. A bilingual concordance tool performs searches on a sentence-aligned parallel corpus. The translator controls the search query and scans the target sentences to locate the translation.



```
... man Blue Angel and the Nordic Swan have led the way in eco-labelling . To scrap these succ ...
... Sweden , as Mrs Paulsen has said , has led the way but there they have looked at , for exa ...
...  European Commission which has actually led the way in ensuring that reformed and transpare ...
... al controls are put in place . We have led the way in reforming those financial controls w ...
...  in the wake of ISA , fish farmers have led the way in adhering to an ISA code . This is p ...
...  and I applaud the fact that the EU has led the way in this area . HIV , tuberculosis and  ...
```

```
De Duitse " Blauwe Engel " en de noordse " Zwaan " hebben de trend gezet op het gebied van milieukeuren .
Zweden heeft , zoals mevrouw Paulsen heeft gezegd , het goede voorbeeld gegeven . Men heeft daar
bijvoorbeeld naar de verhitting van voer gekeken en niet naar de botte methode om dieren af te maken .
Net als in het geval van de Palestijnse Autoriteit , waarnaar hij verwees , was het de Europese Commissie die er
als eerste voor heeft gezorgd dat er hervormde en transparante financiële controles zijn ingevoerd .
Wij hebben het voortouw genomen bij het hervormen van die financiële controles die iedereen nu steunt en
hanteert .
Van Schotland kunnen we misschien nog iets leren : daar hebben de viskwekers naar aanleiding van de
problemen met ISA zelf het voortouw genomen door een ISA-code in te stellen .
Persoonlijk verwelkom ik van harte de inhoud van het document van de Commissie en de voorstellen van de
rapporteur , en ik juich het toe dat de EU op dit terrein een voorbeeld geeft .
```

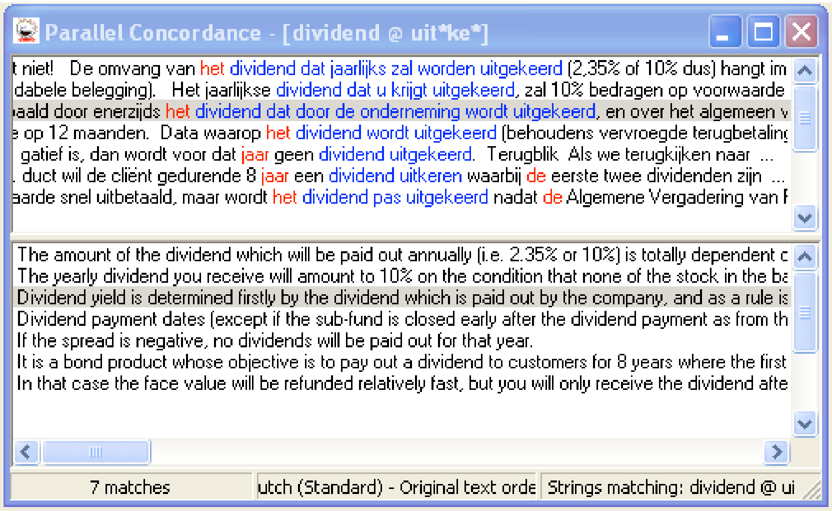
6 matches | nglish (United Kingdom) - Original text orde | Strings matching: led the way

Figure 8.4: Bilingual concordance window in Paraconc with a contiguous search query.

If a bilingual concordance tool is used as a translation aid to solve lexical translation problems, relatively large parallel corpora are needed. A large, freely available parallel corpus is Europarl, which contains parallel texts in eleven European languages (Koehn, 2005). For the language pairs Dutch-English and Dutch-French, the Dutch Parallel Corpus (see section 2.2.3) will be available soon.

Figure 8.4 presents an example of a concordance search for the expression *led the way* in a bilingual corpus. The parallel corpus search offers several Dutch translation suggestions: *de trend zetten*, *als eerste voor iets zorgen*, *het (goede) voorbeeld geven*, *het voortouw nemen*, and the like.

Figure 8.5 presents a concordance search that was performed for the discontiguous expression *dividend...uitkeren*. Paraconc supports wildcards and discontinuity in its search queries, which makes it possible to look for variants of the verb *uitkeren* (*uitgekeerd*, *uitkeert*, and the like) by means of one search query.



t niet!   De omvang van het dividend dat jaarlijks zal worden uitgekeerd (2,35% of 10% dus) hangt im
dabele belegging).   Het jaarlijkse dividend dat u krijgt uitgekeerd, zal 10% bedragen op voorwaarde
aald door enerzijds het dividend dat door de onderneming wordt uitgekeerd, en over het algemeen v
e op 12 maanden. Data waarop het dividend wordt uitgekeerd (behoudens vervroegde terugbetaling
 gatief is, dan wordt voor dat jaar geen dividend uitgekeerd. Terugblik Als we terugkijken naar ...
. duct wil de cliënt gedurende 8 jaar een dividend uitkeren waarbij de eerste twee dividenden zijn ...
aarde snel uitbetaald, maar wordt het dividend pas uitgekeerd nadat de Algemene Vergadering van F

The amount of the dividend which will be paid out annually (i.e. 2.35% or 10%) is totally dependent c
The yearly dividend you receive will amount to 10% on the condition that none of the stock in the ba
Dividend yield is determined firstly by the dividend which is paid out by the company, and as a rule is
Dividend payment dates (except if the sub-fund is closed early after the dividend payment as from th
If the spread is negative, no dividends will be paid out for that year.
It is a bond product whose objective is to pay out a dividend to customers for 8 years where the first
In that case the face value will be refunded relatively fast, but you will only receive the dividend afte

7 matches | utch (Standard) - Original text orde | Strings matching: dividend @ ui

Figure 8.5: Bilingual concordance window in Paraconc with a discontiguous search query.

Bilingual concordance systems cannot be seen as a replacement for translation memory tools. As Bowker and Barlow (2004) conclude, the two technologies may be considered complementary.

8.7 Comparison with our alignment systems

It would have been interesting to plug in our aligned chunks in Similis and evaluate a sub-sentential translation memory system that uses the output of our alignment modules. Unfortunately, Similis does not yet support the import of aligned chunks generated by other systems. Another option – creating a sub-sentential translation memory system from scratch – fell outside the scope of this research project. As we explained above, sub-sentential alignment is just one of the tasks of a sub-sentential translation memory system. A sub-sentential translation memory system must also be able to identify (fuzzy) matches at sub-sentential level and have a scoring and selection mechanism to choose the best match in the case of multiple translation suggestions. Hence, the only meaningful comparison that could be done off-line was comparing the alignment quality of the different systems.

We repeated the test described in section 8.4.1 with two of our sub-sentential alignment systems. The first system is the anchor chunk system that uses the probabilistic dictionary trained on the 9.3M word corpus as lexical resource, which was described in chapter 5.

The second system is the more precise system of the chunk-based extension to the intersected GIZA++ alignments using Log-Likelihood ratio as validation technique, viz. the *10Lex* system. The *10Lex* system uses the extraction corpus of short sentences and only retains lexicalized translation rules or abstract rules containing lexical indices in the validation step (see chapter 6). The results are presented in table 8.11.

Similis

Domain	Correct	Partially	Wrong	Aligned words
Medical	80%	18%	2%	42%
Financial	70%	25%	5%	36%
Journalistic	67%	26%	7%	31%

Anchor Chunk System

Domain	Correct	Partially	Wrong	Aligned words
Medical	94%	6%	0%	38%
Financial	89%	11%	0%	34%
Journalistic	91%	8%	1%	33%

Bootstrapping System

Domain	Correct	Partially	Wrong	Aligned words
Medical	92%	7%	1%	44%
Financial	87%	10%	3%	37%
Journalistic	91%	6%	3%	38%

Table 8.11: Percentage of correctly, partially correct or wrongly aligned chunks and percentage of aligned words in Similis and in our anchor chunk and bootstrapping system.

The *Anchor Chunk* system is the most precise (has the highest percentage of correctly aligned chunks), but is the system with the lowest recall (has the lowest percentage of aligned words). The selected *bootstrapping* system has the highest recall and correctly aligns more words than Similis.

8.8 Summary

We carried out several experiments to assess the usefulness of two different types of translation memory systems (a sentence-based and a sub-sentential translation memory system) on different text types. We extracted three subcorpora of approximately the same size from different text types from the Dutch Parallel Corpus to populate the translation memories. We also extracted three source language texts to be translated.

The analysis functions of both translation memory systems were used to assess the usefulness of the translation memory for the given translation task. We pre-translated the source language documents to be translated and manually inspected all translation suggestions.

On the basis of these experiments we can conclude that sub-sentential translation memory systems are a move in the right direction. Because they look for matches at both the sentential and sub-sentential levels, they cover all functions of sentence-based translation memory systems. Furthermore, they provide useful translation suggestions for terminological units and other fixed expressions. For more flexible expressions (idiomatic expressions and collocations), less automated bilingual concordance programs may be more beneficial.

However, the performance of the sub-sentential TM system that we tested is not yet optimal, as less useful translation suggestions for basic vocabulary words and fuzzy chunk matches often offer translators more distraction than benefit.

In order for sub-sentential translation memories to exploit the full potential of translation memories, better word alignment algorithms are necessary so as to improve both precision (the quality of the chunk alignments) and recall (align more flexible units).

We demonstrated that the different alignment systems that we have developed should be able to generate more aligned chunks of better quality.

CHAPTER 9

Summary and Future Work

9.1 Summary

This final chapter gives an overview of the main achievements of this thesis and offers some prospects for future work.

The main objective of this thesis was the development of a sub-sentential alignment module that could be of value to human translators and to developers of computer-assisted translation tools. The central research question of this thesis was to what extent it would be possible to develop a robust automatic sub-sentential alignment module that aligns segments with a very high precision.

As in most language technology projects, our research plan comprised different steps: from data collection, preprocessing and the creation of an evaluation framework over the actual development of the sub-sentential alignment system to the validation of the alignment system in real-life applications.

Sentence-aligned parallel corpora formed an indispensable resource for this research project. To cover a wide range of syntactic and stylistic phenomena that emerge from different writing and translation styles, we used parallel texts of different text types. The most important data sources we used were the Europarl corpus (Koehn 2005) and the Dutch Parallel Corpus (Macken et al. 2007).

The Europarl corpus contains the proceedings of the European Parliament plenary debates in each official language of the European Union. The texts are a written reproduction of prepared speeches and cover a wide range of subject fields. As we preferred to work with direct translations of original English source sentences produced by native English speakers into Dutch, we made use of the language tag and the speaker identity tag to select those interventions.

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. As partner of the core research team, we were closely involved in all processing steps of the DPC project. In chapter 2, we described the sentence alignment process that was adopted for the DPC project. We discussed in detail how the output of different alignment tools was combined to minimize the human effort that is needed to achieve nearly 100% alignment accuracy at sentence level.

The sub-sentential alignment system we developed is chunk-driven and requires shallow linguistic processing tools for the source and target language, i.e. part-of-speech taggers and chunkers. A lemmatizer was used to abstract over word forms before dictionary look-up. Several publicly available linguistic processing tools were used: we used the combined memory-based PoS tagger/lemmatizer for English (Daelemans and van den Bosch 2005), TADPOLE for Dutch (van den Bosch, Busser, Daelemans and Canisius 2007) and Treetagger for French, Italian and English (Schmid 1994) in the terminology extraction sub-project. As each language has its own part-of-speech tag set, mapping tables were defined in order to compare the different part-of-speech codes efficiently.

We opted for shallow parsing – as opposed to full parsing – to retrieve syntactic information from the sentences. Rule-based chunkers for Dutch and English were developed in the framework of this research project. A language-independent variant of the chunkers was developed for the terminology sub-project described in chapter 7.

Next to sentence-aligned parallel texts enriched with different levels of linguistic analysis (part-of-speech codes, lemmas and chunk boundaries), our sub-sentential alignment system makes use of a bilingual dictionary to retrieve lexical correspondences. We experimented with two different types of bilingual dictionaries: a handcrafted bilingual dictionary and probabilistic bilingual dictionaries that were derived from a parallel corpus by using a statistical word alignment package. The handcrafted bilingual lexicon was extracted from the English-Dutch and Dutch-English NL-Translex lexicons (Goetschalckx et al. 2001).

Experiments were carried out with two types of statistical word alignment packages. In chapter 5, the probabilistic bilingual dictionaries were derived from the IBM Model One output on different variants of the Europarl corpus (a 2.4, 5.6, and 9.3 million word sub-corpus). We made use of the Perl implementation of IBM Model One, which is part of the Microsoft Bilingual Sentence Aligner (Moore 2002). In chapter 6, the system does not start from a bilingual dictionary but builds directly further on the output of GIZA++ (Och and Ney 2003), a state-of-the-art tool for statistical word alignment that implements the IBM models 1–5 (Brown et al. 1993) and that is computationally more complex than IBM model One.

The sub-sentential alignment system we have built is conceived as a cascaded model consisting of two phases. In the first phase, anchor chunks are retrieved, i.e. chunks that can be linked with a very high precision on the basis of lexical correspondences and syntactic similarity. In the second phase, we focussed on the more complex translational correspondences on the basis of observed translation shift patterns. The anchor chunks of the first phase were used to limit the search space in the second phase.

In chapter 5, the performance of a system using a handcrafted bilingual dictionary was compared with a system using a probabilistic bilingual dictionary derived from the IBM Model One alignments. We demonstrated that although the handcrafted dictionary was twice the size of the probabilistic dictionary, the obtained recall scores were lower. No difference in precision was observed for the retrieved anchor chunks.

In chapter 6, the anchor chunks were built on the basis of the intersected GIZA++ word alignments. The intersected alignment points were very precise and had a higher recall than the IBM Model One word alignments. The anchor chunks were used in a bootstrapping approach to retrieve the more complex translational correspondences. We developed three different methods to validate the extracted candidate rules in the bootstrapping process. The first method makes use of a manually created reference corpus. The second one uses manually created filtering heuristics, whereas the third method is fully automatic and filters on the basis of the Log-Likelihood ratio.

An important aspect of all language technology projects is evaluation. In order to be able to compare different alignment systems, some objective means of evaluating their performance is required. Therefore, we created an annotation scheme and a Gold Standard for Dutch-English translational correspondence. 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. A multi-level annotation was introduced in the case of divergent translations: fuzzy links were used to connect paraphrased sections, regular links to connect corresponding words within the paraphrased sections. As the evaluation of word or chunk alignment systems is not a trivial task, different metrics were used to compare the output of the automatic alignment systems with the manually created Gold Standard.

Extensive evaluations were carried out on the two phases of our sub-sentential alignment system. The objective of the first phase was to link *anchor chunks*, i.e. chunks that can be linked with a very high precision. In the anchor chunk system described in chapter 5, on average 45-60% of the words could be linked with a precision ranging from 90% to 98%. In chapter 6, it was demonstrated that the proposed bootstrapping systems improved the recall of the intersected GIZA++ word alignments without sacrificing precision, which makes the resulting alignments useful for incorporation in CAT-tools or bilingual terminology extraction tools.

We validated the sub-sentential alignment system in two different applications. In chapter 7, we described how the anchor chunk system was used to guide bilingual terminology extraction. We also demonstrated the language independence of our sub-sentential alignment approach by carrying out terminology extraction experiments on three different language pairs, viz. French-English, French-Italian and French-Dutch.

In chapter 8, we examined the usefulness of sub-sentential translation memory systems by comparing a sentence-based translation memory system (Trados Translator's workbench) with a sub-sentential translation memory system (Similis). We demonstrated that sub-sentential translation memory systems provide useful sub-sentential translation suggestions for terminological units and other fixed expressions. We compared the alignment quality of our sub-sentential alignment system with Similis and demonstrated that both the anchor chunk system described in chapter 5 and the bootstrapping system described in chapter 6 should be able to generate more and better aligned chunks than Similis.

9.2 Future Work

The sub-sentential alignment system we developed is far from perfect. There are at least two areas for improvement.

The first area deals with Dutch compounds. In Dutch, compounds are often not separated by means of white space characters and hence form a single word. In chapter 7, we carried out preliminary experiments by adding a decompounding module as a fall-back solution. When no lexical correspondences could be retrieved from the bilingual dictionary, a second dictionary look-up was performed for the compound parts. However, as the bilingual dictionary was derived from a corpus of word forms, the dictionary only contained Dutch compound parts that were also present in the corpus as separate words. We expect that splitting the compounds in the Dutch part of the parallel corpus prior to statistical word alignment will have a larger positive impact on the word alignment results.

The second area deals with multiword expressions of different kinds. Some structures were only partially linked by our sub-sentential alignment module. Adding a module that can detect multiword expressions (e.g. Dutch separable verbs, English phrasal verbs, multiword prepositions, and the like) would allow us to group the detected multiword parts together before chunk alignment.

On the application side, we would like to incorporate our sub-sentential alignment system in an open source translation memory system (OmegaT[1] or Anaphraseus[2]). Care should be taken however. We demonstrated that Similis also offers sub-sentential translation suggestions that are of no use to human translators (e.g. basic vocubulary words). A better understanding of the translation problems that human translators encounter is required in order to develop CAT-systems that can accommodate for the human translator's needs.

For terminology extraction, the filtering statistics need more attention. We plan to create Gold Standards for bilingual terminology extraction, which will allow us to set the threshold filtering parameters in an automatic way and to compute recall next to precision.

An application domain that was not covered in this thesis due to time constraints was Machine Translation. In the domain of Machine Translation, the currently best performing statistical systems are based on phrase-based models (Koehn:2009), which in fact assemble translations of different sub-sentential units. As our chunk-based extension to the intersected GIZA++ alignments aligns chunks rather than words and is able to align discontinuous chunks, we hope that the incorporation of these precise chunks in the SMT phrase table has a positive impact on Machine Translation quality as well. As most phrase-based SMT systems do not support discontinuous phrases, important changes in the translation component of the MT systems have to be made.

And last but not least, we would like to share our results with the scientific community and plan to release a clean version of our Java code as open source.

[1] http://www.omegat.org/en/downloads.html

[2] http://anaphraseus.sourceforge.net/

APPENDIX A

Mapping table for the English and Dutch PoS codes

The English and Dutch part-of-speech taggers use different PoS tag sets. The English PoS tagger uses the Penn Treebank tag set (Marcus et al. 1993), which is a coarse-grained tag set of 45 distinct tags. The Dutch PoS tagger uses the CGN PoS tag set (Van Eynde et al. 2000), which codes a wide range of morphosyntactic features (e.g. singular, plural, case information, tense) as attributes to the word class. In total, 316 distinct full tags are discerned.

To be able to compare the English and Dutch part-of-speech codes efficiently, the English and Dutch PoS codes are mapped to a newly defined coarse-grained tag set. To map the English tags, the full English tag is mapped to the new tag. One exception is handled by means of a rules: as the English PoS tag "IN" is used for both prepositions and subordinate conjunctions, this tag was further disambiguated during chunking.

To map the Dutch tags, regular expressions are used to map the main part of the Dutch tag together with some attributes to the new tag.

Word Class	English Pos	Mapped PoS
opening quote	“	PCT
closing quote	”	PCT
opening parenthesis	(	PCT
closing parenthesis	)	PCT
comma	,	PCT
dash	–	PCT
colon or ellipsis	:	PCT
coordinating conjunction	CC	CONJ-coord
cardinal number	CD	NUM
determiner	DT	DET
existential there	EX	EX
foreign word	FW	FW
preposition or subordinate conjunction	IN	PREP/CONJ-subord
adjective or ordinal numeral	JJ	ADJ
list item marker	LS	PCT
modal auxiliary	MD	V-fin
common noun, sg or mass	NN	N
plural common noun	NNS	N
proper noun sg	NNP	N-prop
plural proper noun	NNPS	N-prop
personal pronoun	PRP	PRON-per
possessive pronoun	PRP$	PRON-pos
relative pronoun	WDT	PRON-rel
interrogative pronoun	WP	PRON-int
pronoun (other)	WP$	PRON
adverb	RB	ADV
wh-adverb	WRB	ADV
particle	RP	PART
symbol	SYM	SYM
preposition/infinitive marker	TO	PREP
interjection	UH	INT
infinitive	VB	V-inf
verb, past tense	VBD	V-fin
verb, present particle/gerund	VBG	V-prpa
verb, past participle	VBN	V-papa
verb, present tense (not 3p)	VBP	V-fin
verb, present tense (3p)	VBZ	V-fin

Table A.1: English PoS mapping table.

Word Class	Dutch Pos	Mapped PoS
opening quote	LET	PCT
closing quote	LET	PCT
opening parenthesis	LET	PCT
closing parenthesis	LET	PCT
comma	LET	PCT
dash	LET	PCT
colon or ellipsis	LET	PCT
coordinating conjunction	VG(neven	CONJ-coord
cardinal number	TW(hoofd	NUM
ordinal number	TW(rang	NUM
determiner	LID	DET
foreign word	SPEC(vreemd	FW
subordinate conjunction	VG(onder	CONJ-subord
adjective or ordinal numeral	ADJ	ADJ
ordinal number, prenominal	TW(rang,prenom	ADJ
cardinal number, prenominal	TW(hoofd,prenom	ADJ
past participle, prenominal	WW(vd,prenom	ADJ
present participle, prenominal	WW(od,prenom	ADJ
list item marker	SPEC	PCT
common noun, sg	N(soort, ev	N
plural common noun	N(soort,mv	N
proper noun sg	N(eigen,ev	N-prop
plural proper noun	N(eigen,mv	N-prop
adjective, nominalised	ADJ(nom,	N
cardinal number, nominalised	TW(hoofd,nom	N
ordinal number, nominalised	TW(rang,nom	N
past participle, nominalised	WW(vd,nom	N
present participle, nomimalised	WW(od,nom	N
genitive marker	N(*,gen)	N-gen
personal pronoun	VNW(pers	PRON-per
possessive pronoun	VNW(bez	PRON-pos
indefinitive pronoun	VNW(onbep	PRON-ind
demonstrative pronoun	VNW(aanw	PRON-dem
relative pronoun	VNW(betr	PRON-rel
interrogative pronoun	VNW(vb	PRON-int
pronoun (other)	VNW	PRON

Word Class	Dutch Pos	Mapped PoS
adverb	BW	ADV
wh-adverb	BW	ADV
symbol	SPEC(sym	SYM
preposition	VZ	PREP
interjection	TSW	INT
infinitive	WW(inf	V-inf
verb, past tense	WW(pv,verl	V-fin
verb, present particle	WW(od	V-prpa
verb, past participle	WW(vd	V-papa
verb, present tense	WW(pv,tgw	V-fin

Table A.2: Dutch PoS mapping table.

APPENDIX B

Annotation Guidelines

B.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 2001c, 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 2001a, Merkel 1999a, Och and Ney 2000, 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 1999a) 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.

B.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[2], and vice versa.

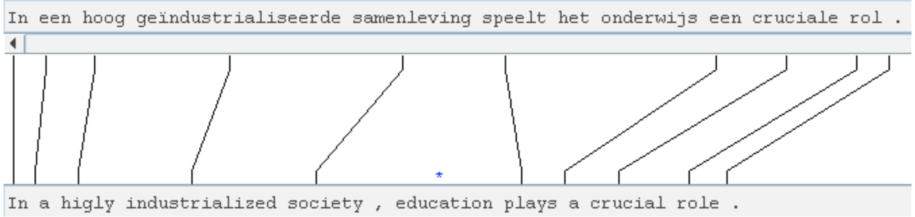
[1] http://www.cs.utah.edu/∼hal/HandAlign/

[2] 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

To determine this minimal language unit, two major rules can be formulated (Merkel 1999a, 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

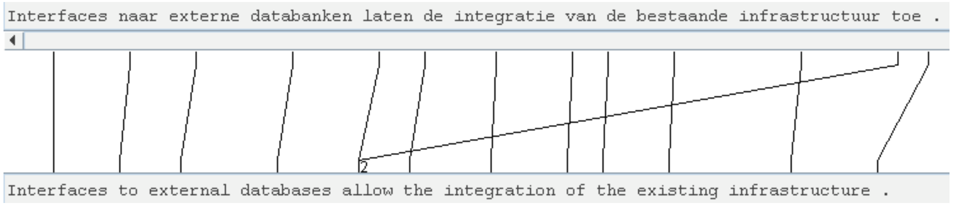
In the first example there is word-by-word correspondence for all words except for *het onderwijs* ∼ *education*: in the English sentence there is no definite article.



```
In een hoog geïndustrialiseerde samenleving speelt het onderwijs een cruciale rol .
```

```
In a higly industrialized society , education plays a crucial role .
```

The corresponding units are not necessarily contiguous[3], e.g. *laten...toe* ∼ *allow.*



```
Interfaces naar externe databanken laten de integratie van de bestaande infrastructuur toe .
```

```
Interfaces to external databases allow the integration of the existing infrastructure .
```

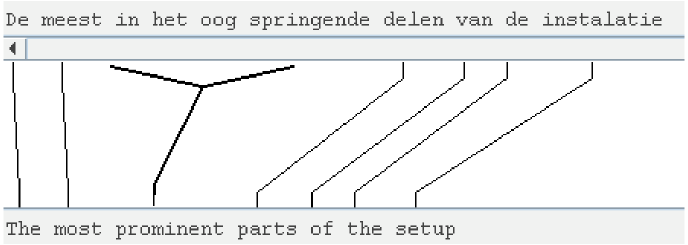
In most translations however, translational correspondences are more complex, and only for some words, word-by-word correspondences can be found. The rest of the sentence is translated on the level of combination of words.

elements, taken separately, do not correspond to equivalents in the translated text.

[3] 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.

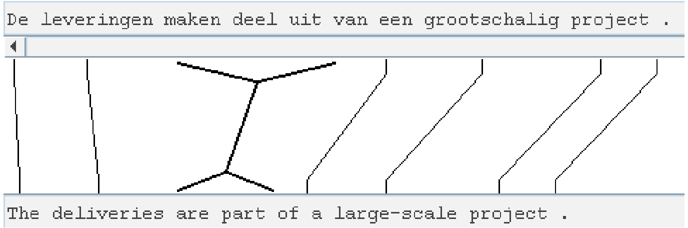
Phraseological units

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*.



De meest in het oog springende delen van de instalatie

The most prominent parts of the setup



De leveringen maken deel uit van een grootschalig project .

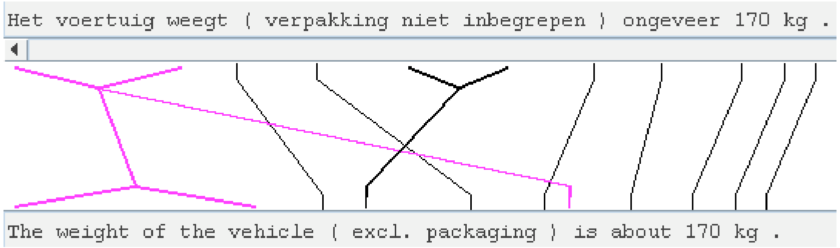
The deliveries are part of a large-scale project .

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).

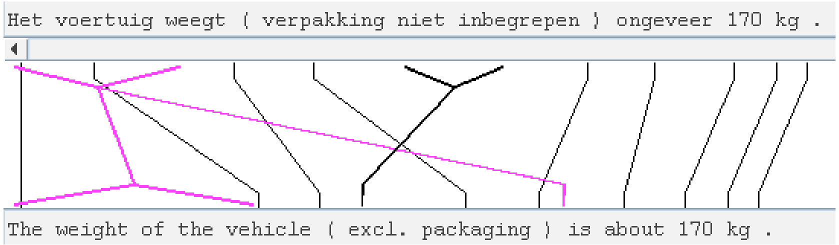
(1) Fuzzy link: *het voertuig weegt* ∼ *the weight of the vehicle … is*



Het voertuig weegt (verpakking niet inbegrepen) ongeveer 170 kg .

The weight of the vehicle (excl. packaging) is about 170 kg .

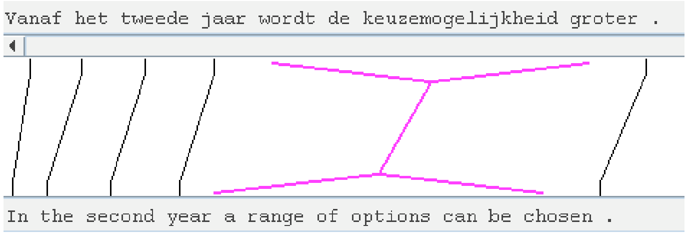
(2) Regular links: *Het* ∼ *the*, *voertuig* ∼ *vehicle*



Het voertuig weegt (verpakking niet inbegrepen) ongeveer 170 kg .

The weight of the vehicle (excl. packaging) is about 170 kg .

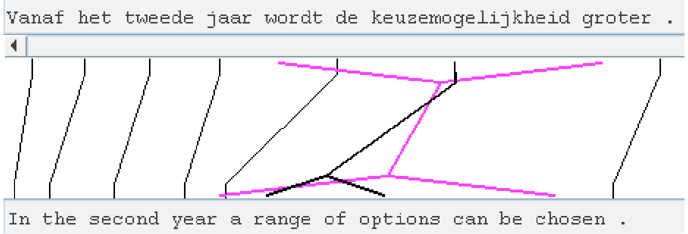
(1) Fuzzy link: *wordt de keuzemogelijkheid groter* ∼ *a range of options can be chosen*



Vanaf het tweede jaar wordt de keuzemogelijkheid groter .

In the second year a range of options can be chosen .

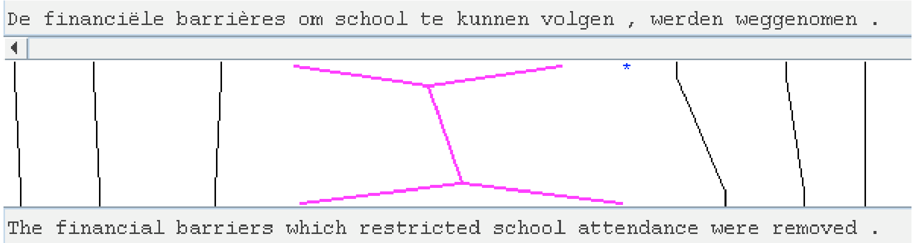
(2) Regular links: *de* ∼ *a*, *keuzemogelijkheid* ∼ *range of options*



Vanaf het tweede jaar wordt de keuzemogelijkheid groter .

In the second year a range of options can be chosen .

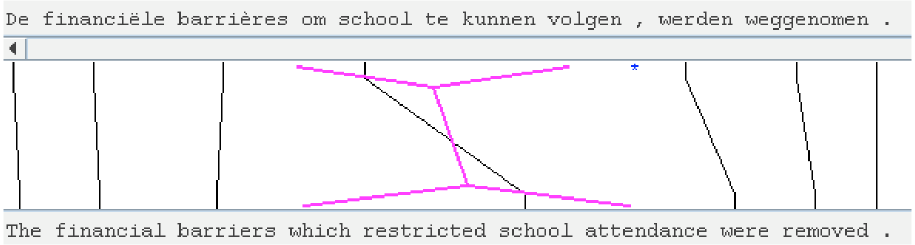
(1) Fuzzy link: *om school te kunnen volgen* ∼ *which restricted school attendance*



De financiële barrières om school te kunnen volgen , werden weggenomen .

The financial barriers which restricted school attendance were removed .

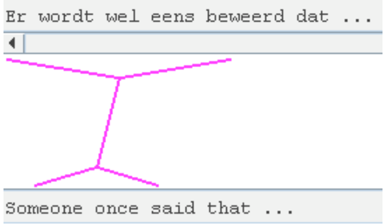
(2) Regular link: school ∼ school



De financiële barrières om school te kunnen volgen , werden weggenomen .

The financial barriers which restricted school attendance were removed .

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

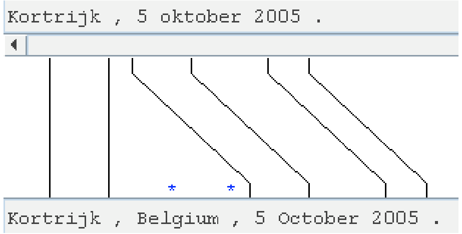


Er wordt wel eens beweerd dat ...

Someone once said that ...

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.



Kortrijk , 5 oktober 2005 .

Kortrijk , Belgium , 5 October 2005 .

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.

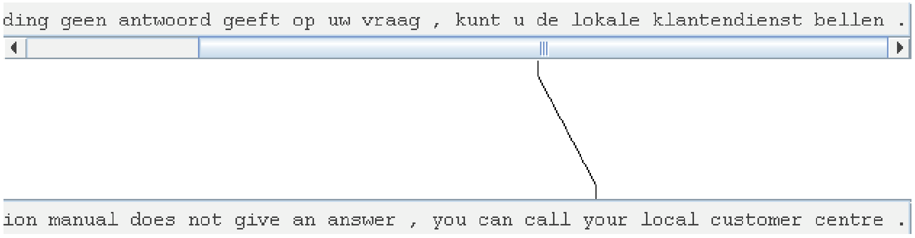
The multi-level annotation scheme is only used for regular links within fuzzy links. Do not use regular links within regular links.

B.3 Detailed Guidelines

B.3.1 Noun Phrases

Determiners

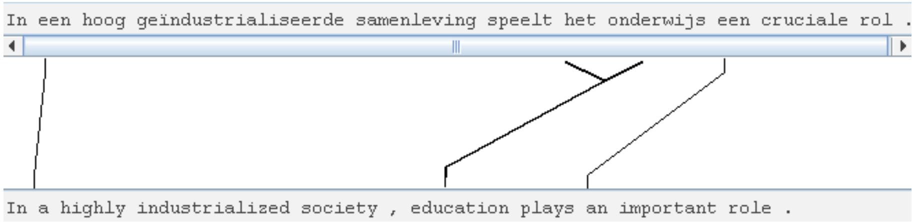
Determiners can be connected with a regular link, regardless whether they are articles or possessive pronouns.



ding geen antwoord geeft op uw vraag , kunt u de lokale klantendienst bellen .

ion manual does not give an answer , you can call your local customer centre .

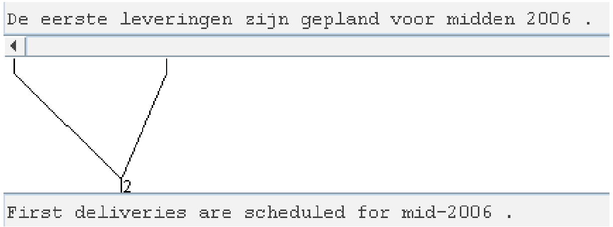
Extra determiners in source or target language should be linked together with their noun to the noun's translation with a regular link.



In een hoog geïndustrialiseerde samenleving speelt het onderwijs een cruciale rol .

In a highly industrialized society , education plays an important role .

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).



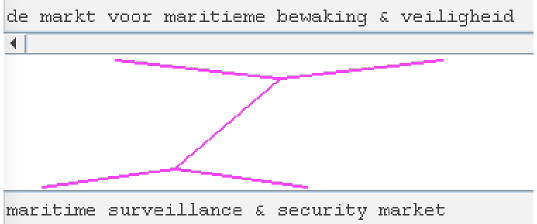
De eerste leveringen zijn gepland voor midden 2006 .

First deliveries are scheduled for mid-2006 .

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).

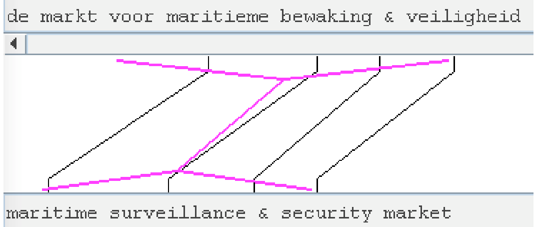
(1) Fuzzy link: *voor maritieme bewaking & veiligheid* ∼ *maritime surveillance & security*



de markt voor maritieme bewaking & veiligheid

maritime surveillance & security market

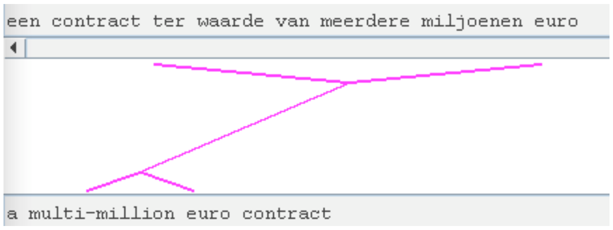
(2) Regular links: *maritieme* ∼ *maritime*, *bewaking* ∼ *surveillance*, *&* ∼ *&*, *veiligheid* ∼ *security*



de markt voor maritieme bewaking & veiligheid

maritime surveillance & security market

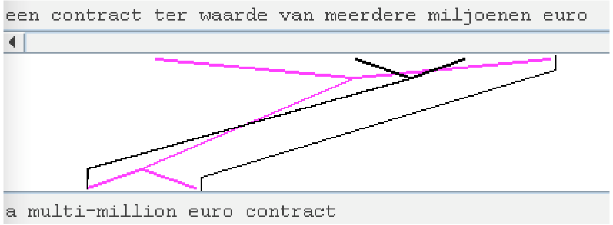
(1) Fuzzy link: *ter waarde van meerdere miljoenen euro* ∼ *multi-million euro*



een contract ter waarde van meerdere miljoenen euro

a multi-million euro contract

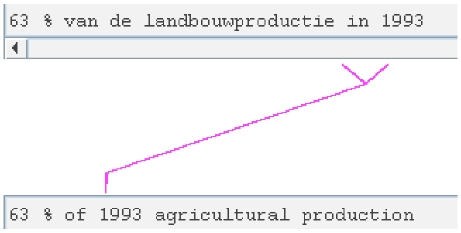
(2) Regular links: *meerdere miljoenen* ∼ *multi-million*, *euro* ∼ *euro*



een contract ter waarde van meerdere miljoenen euro

a multi-million euro contract

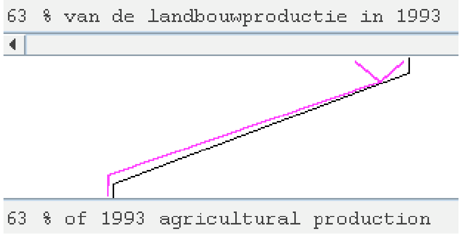
(1) Fuzzy link: *in 1993* ∼ *1993*



63 % van de landbouwproductie in 1993

63 % of 1993 agricultural production

(2) Regular link: *1993* ∼ *1993*

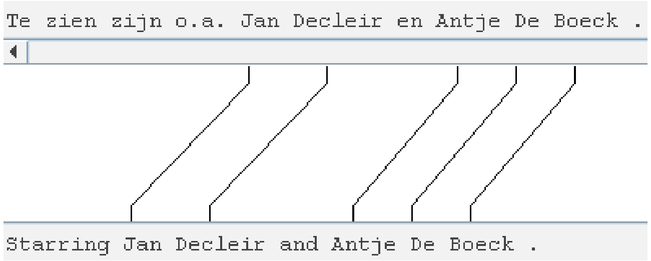


63 % van de landbouwproductie in 1993

63 % of 1993 agricultural production

Proper names

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

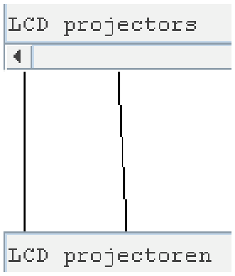


Te zien zijn o.a. Jan Decleir en Antje De Boeck .

Starring Jan Decleir and Antje De Boeck .

Compounds

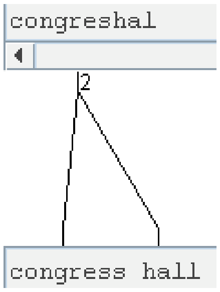
Link the corresponding subparts of the compounds by means of a regular link.



LCD projectors

LCD projectoren

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



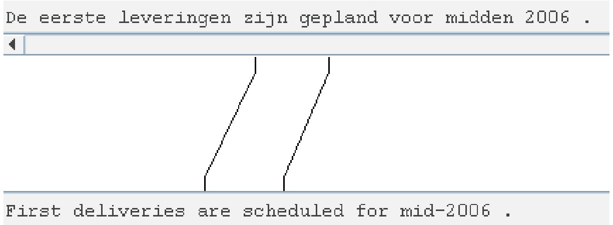
congreshal

congress hall

B.3.2 Verb Phrases

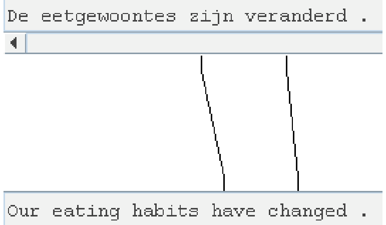
Auxiliary verbs

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



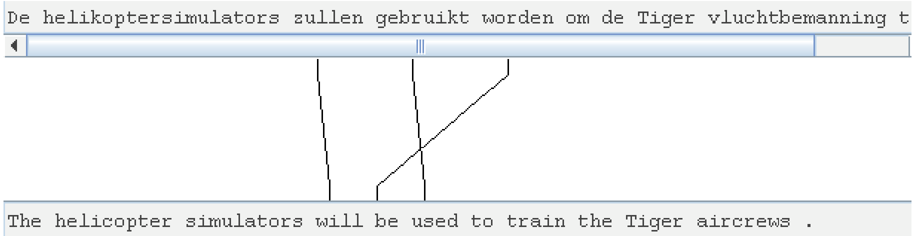
De eerste leveringen zijn gepland voor midden 2006 .

First deliveries are scheduled for mid-2006 .



De eetgewoontes zijn veranderd .

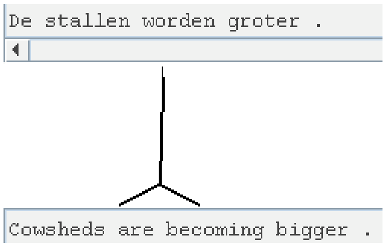
Our eating habits have changed .



De helikoptersimulators zullen gebruikt worden om de Tiger vluchtbemanning t

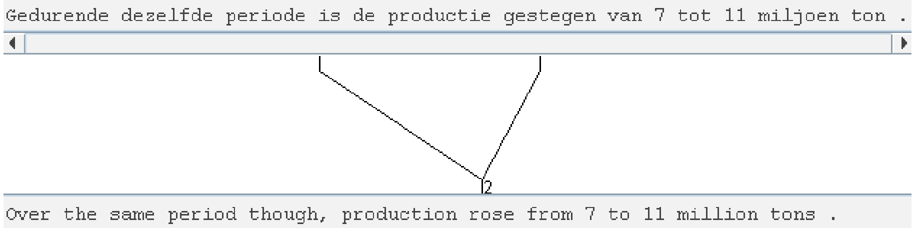
The helicopter simulators will be used to train the Tiger aircrews .

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").



De stallen worden groter .

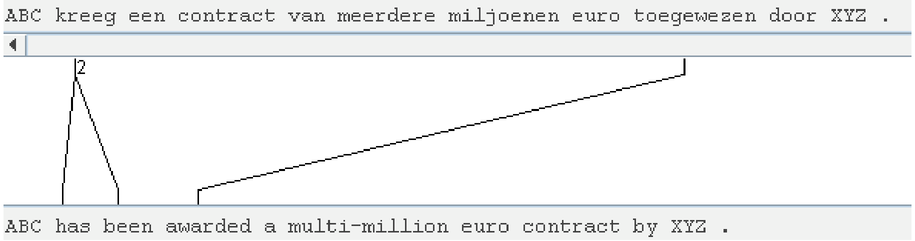
Cowsheds are becoming bigger .



Gedurende dezelfde periode is de productie gestegen van 7 tot 11 miljoen ton .

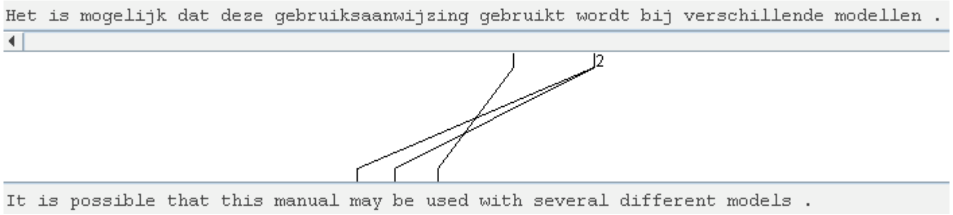
Over the same period though, production rose from 7 to 11 million tons .

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.



ABC kreeg een contract van meerdere miljoenen euro toegewezen door XYZ .

ABC has been awarded a multi-million euro contract by XYZ .

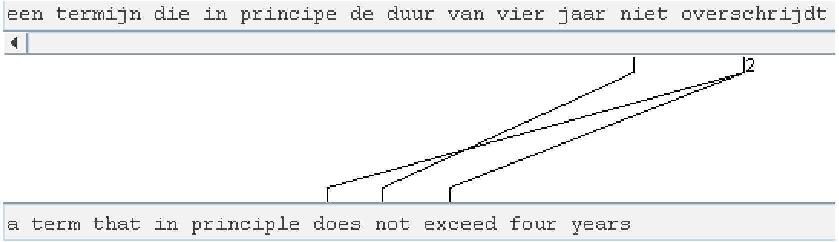


Het is mogelijk dat deze gebruiksaanwijzing gebruikt wordt bij verschillende modellen .

It is possible that this manual may be used with several different models .

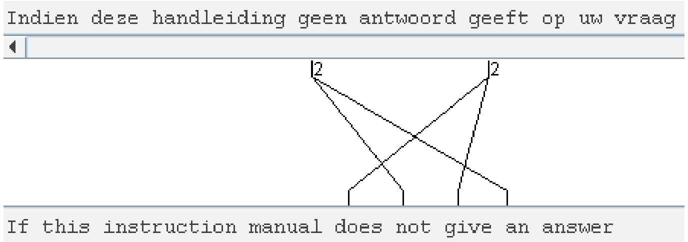
Negation and do-support

Link the auxiliary “do” together with the main verb to the verb’s translation with a regular link.



een termijn die in principe de duur van vier jaar niet overschrijdt

a term that in principe does not exceed four years



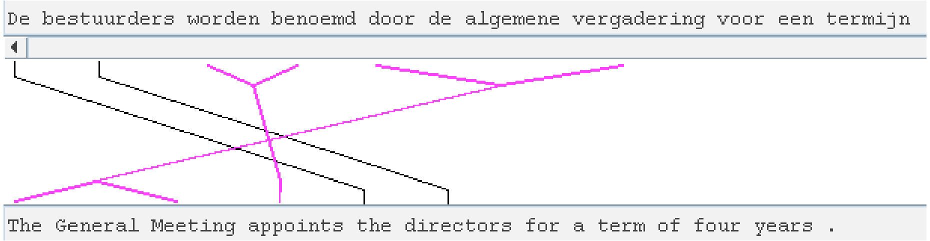
Indien deze handleiding geen antwoord geeft op uw vraag

If this instruction manual does not give an answer

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).

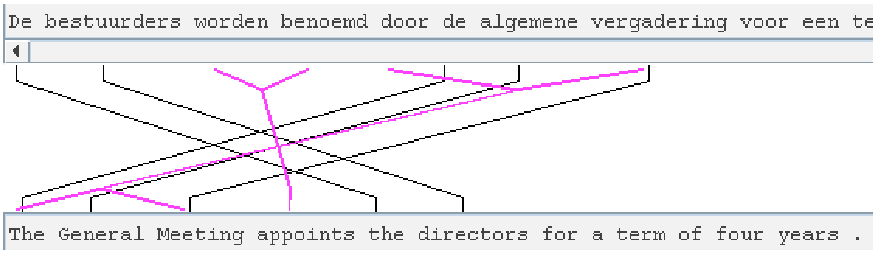
(1) Fuzzy links: *worden benoemd* ∼ *appoints*, *door de algemene vergadering* ∼ *the General Meeting*



De bestuurders worden benoemd door de algemene vergadering voor een termijn

The General Meeting appoints the directors for a term of four years .

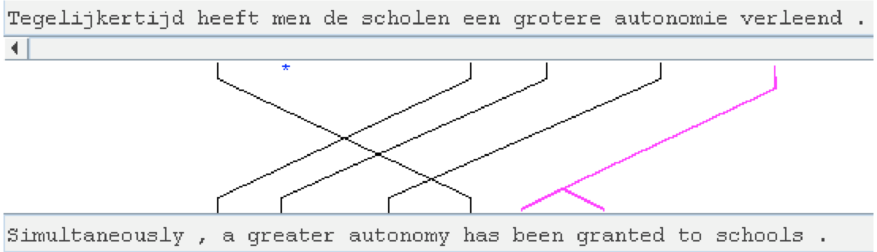
(2) Regular links: *de* ∼ *the*, *algemene* ∼ *General*, *vergadering* ∼ *Meeting*



De bestuurders worden benoemd door de algemene vergadering voor een te

The General Meeting appoints the directors for a term of four years .

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

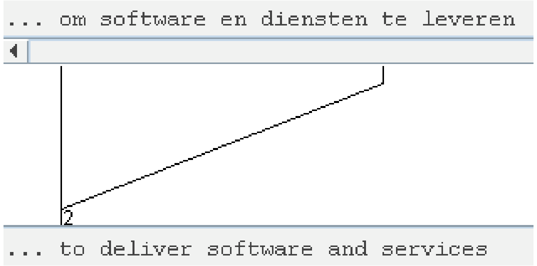


Tegelijkertijd heeft men de scholen een grotere autonomie verleend .

Simultaneously , a greater autonomy has been granted to schools .

Infinitive marker "te"

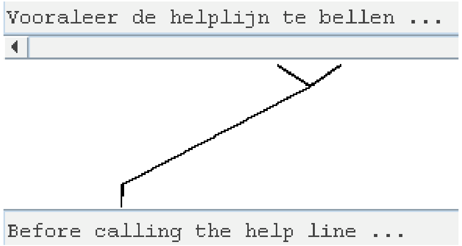
If the Dutch construction "om ... te" corresponds with English "to" use a regular link to connect "om ... te" to "to".



... om software en diensten te leveren

... to deliver software and services

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

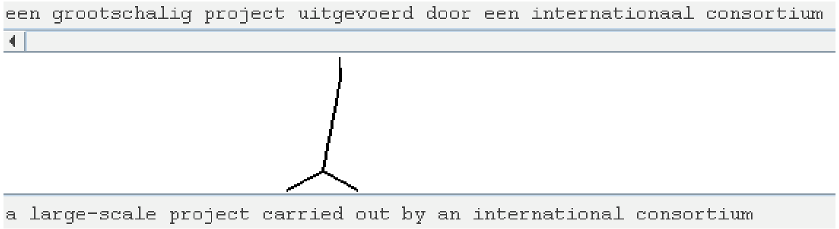


Vooraleer de helplijn te bellen ...

Before calling the help line ...

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.

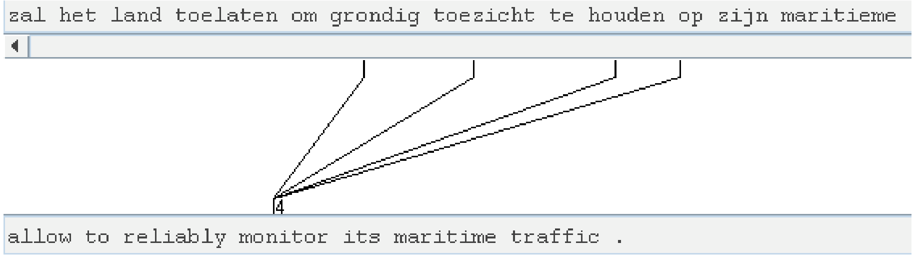


een grootschalig project uitgevoerd door een internationaal consortium

a large-scale project carried out by an international consortium

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.



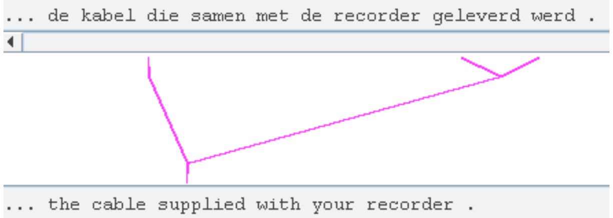
zal het land toelaten om grondig toezicht te houden op zijn maritieme

allow to reliably monitor its maritime traffic .

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).

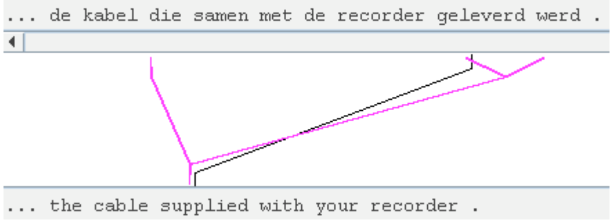
(1) Fuzzy link: *die ... geleverd werd* ∼ *supplied*



... de kabel die samen met de recorder geleverd werd .

... the cable supplied with your recorder .

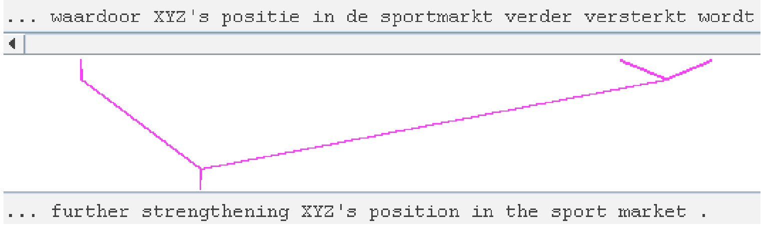
(2) Regular link: *geleverd* ∼ *supplied*



... de kabel die samen met de recorder geleverd werd .

... the cable supplied with your recorder .

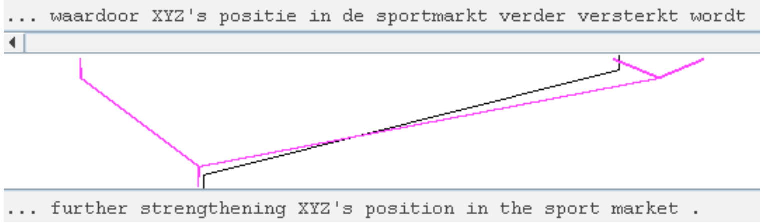
(1) Fuzzy link: *waardoor … versterkt wordt* ∼ *strengthening*



```
... waardoor XYZ's positie in de sportmarkt verder versterkt wordt
```

```
... further strengthening XYZ's position in the sport market .
```

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



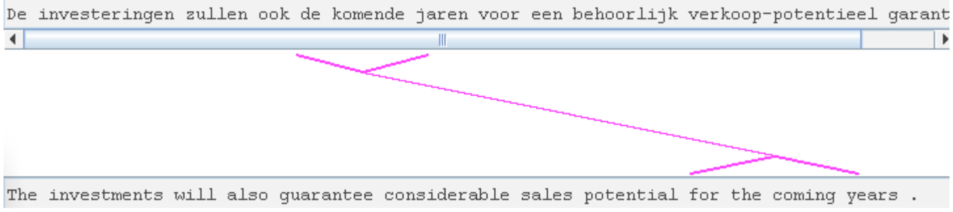
```
... waardoor XYZ's positie in de sportmarkt verder versterkt wordt
```

```
... further strengthening XYZ's position in the sport market .
```

B.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.

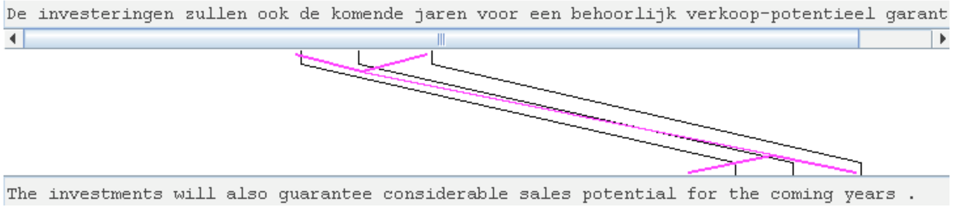
(1) Fuzzy link: *de komende jaren* ∼ *for the coming years*



```
De investeringen zullen ook de komende jaren voor een behoorlijk verkoop-potentieel garant
```

```
The investments will also guarantee considerable sales potential for the coming years .
```

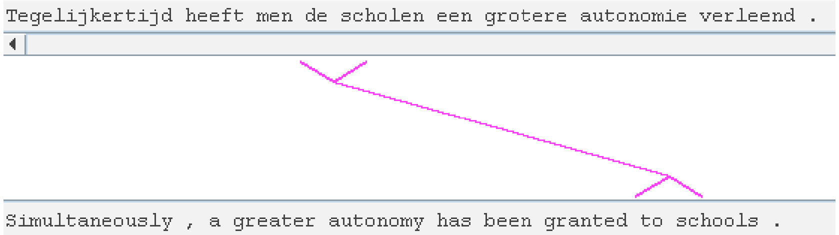
(2) Regular links: *de* ∼ *the*, *komende* ∼ *coming*, *jaren* ∼ *years*



De investeringen zullen ook de komende jaren voor een behoorlijk verkoop-potentieel garant

The investments will also guarantee considerable sales potential for the coming years .

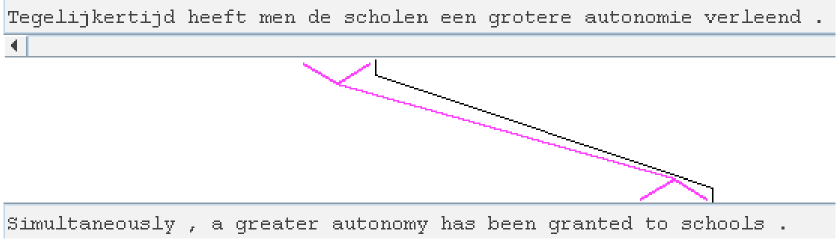
(1) Fuzzy link: *de scholen* ∼ *to schools*



Tegelijkertijd heeft men de scholen een grotere autonomie verleend .

Simultaneously , a greater autonomy has been granted to schools .

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

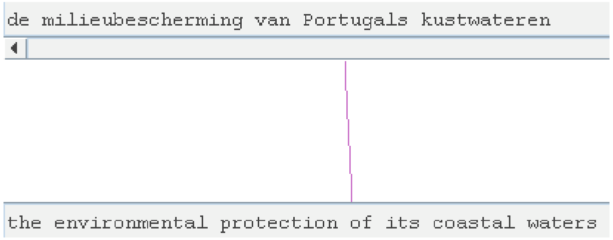


Tegelijkertijd heeft men de scholen een grotere autonomie verleend .

Simultaneously , a greater autonomy has been granted to schools .

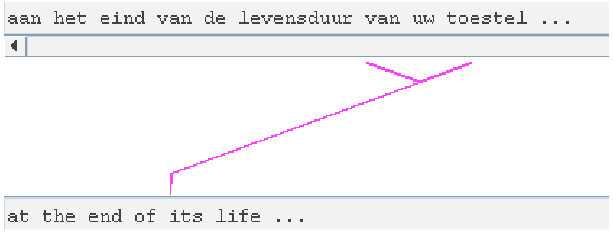
B.3.4 Referring expressions

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



de milieubescherming van Portugals kustwateren

the environmental protection of its coastal waters

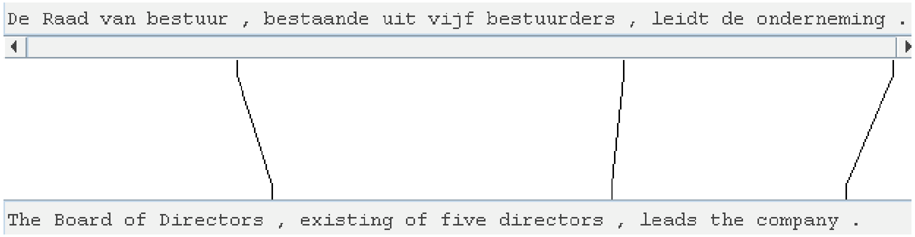


aan het eind van de levensduur van uw toestel ...

at the end of its life ...

B.3.5 Punctuation

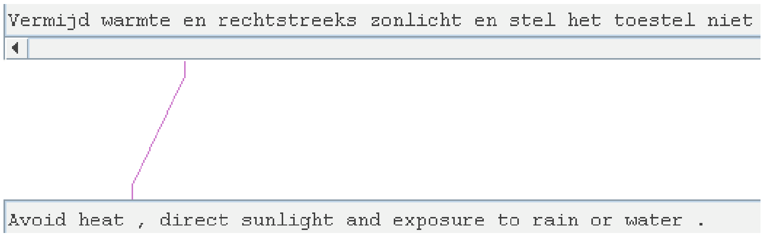
Connect corresponding punctuation marks with a regular link.



De Raad van bestuur , bestaande uit vijf bestuurders , leidt de onderneming .

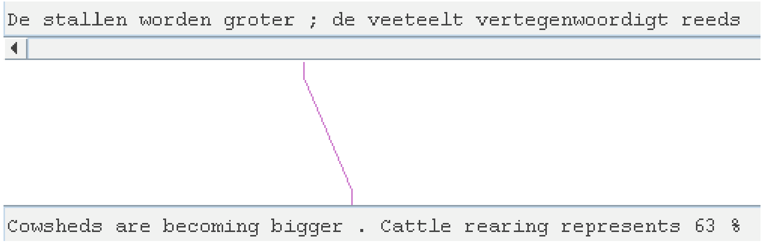
The Board of Directors , existing of five directors , leads the company .

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



Vermijd warmte en rechtstreeks zonlicht en stel het toestel niet

Avoid heat , direct sunlight and exposure to rain or water .

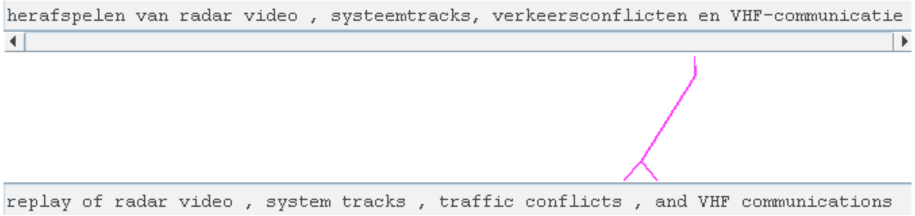


De stallen worden groter ; de veeteelt vertegenwoordigt reeds

Cowsheds are becoming bigger . Cattle rearing represents 63 %

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).

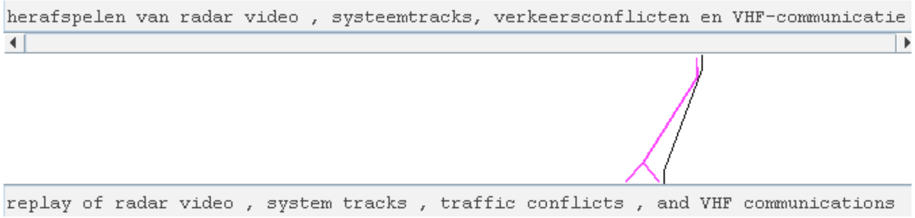
(1) Fuzzy link: *en* ∼ *, and*



herafspelen van radar video , systeemtracks, verkeersconflicten en VHF-communicatie

replay of radar video , system tracks , traffic conflicts , and VHF communications

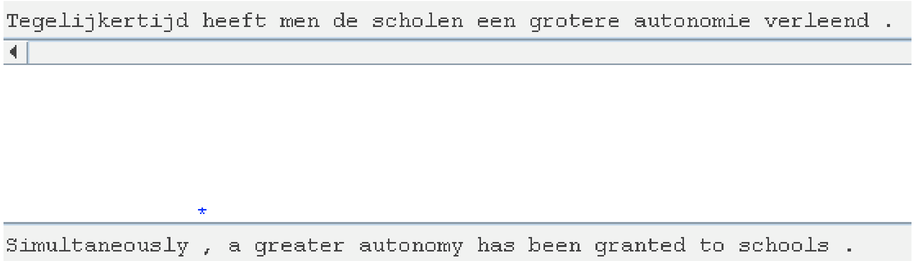
(2) Regular link: *en* ∼ *and*



herafspelen van radar video , systeemtracks, verkeersconflicten en VHF-communicatie

replay of radar video , system tracks , traffic conflicts , and VHF communications

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

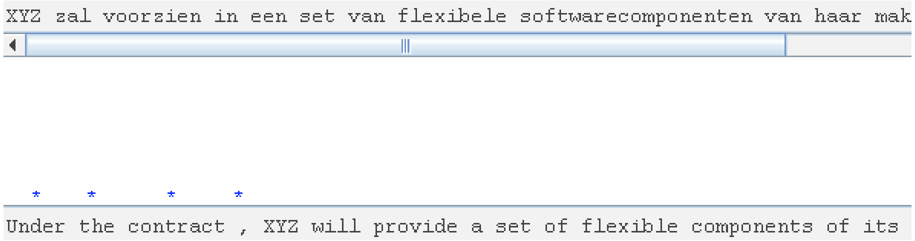


Tegelijkertijd heeft men de scholen een grotere autonomie verleend .

Simultaneously , a greater autonomy has been granted to schools .

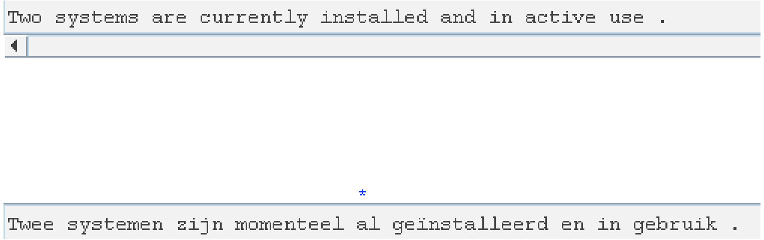
B.3.6 Omissions

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.



XYZ zal voorzien in een set van flexibele softwarecomponenten van haar mak

Under the contract , XYZ will provide a set of flexible components of its

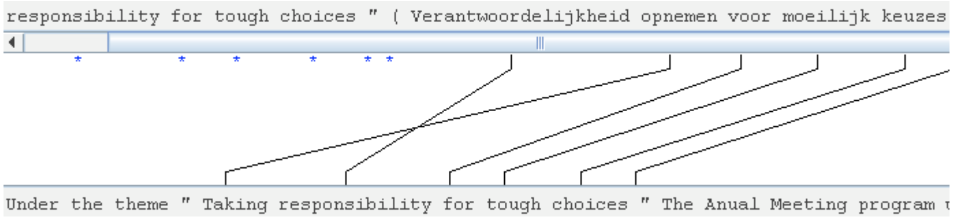


Two systems are currently installed and in active use .

Twee systemen zijn momenteel al geïnstalleerd en in gebruik .

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.



responsibility for tough choices " (Verantwoordelijkheid opnemen voor moeilijk keuzes

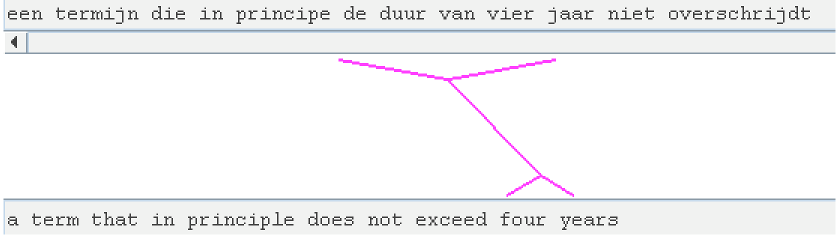
Under the theme " Taking responsibility for tough choices " The Anual Meeting program u

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).

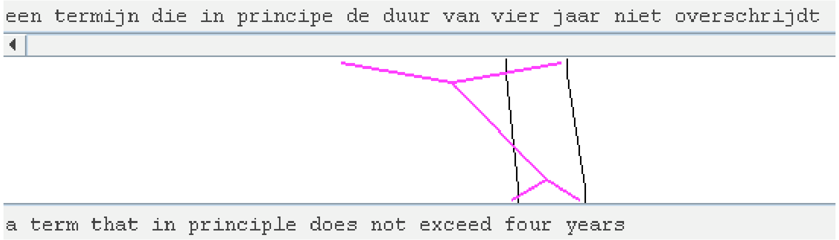
(1) Fuzzy link: *de duur van vier jaar* ∼ *four years*



een termijn die in principe de duur van vier jaar niet overschrijdt

a term that in principle does not exceed four years

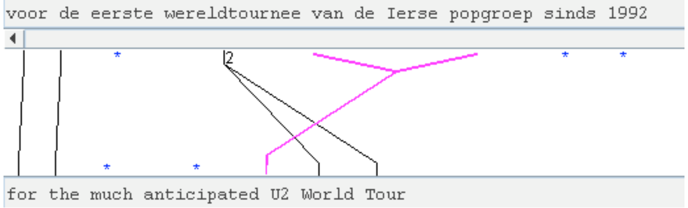
(2) Regular link: *vier jaar* ∼ *four years*



een termijn die in principe de duur van vier jaar niet overschrijdt

a term that in principle does not exceed 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.



voor de eerste wereldtournee van de Ierse popgroep sinds 1992

for the much anticipated U2 World Tour

B.4 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

APPENDIX C

Java classes and methods

The sub-sentential alignment module was implemented in Java 5.0. An object-oriented programming language such as Java groups data and operations (methods) into units called *objects*. Objects are particular instances of a *class*, which defines the abstract characteristics of an object. In an object-oriented programming language, it is possible to create rather intuitive classes.

C.1 Anchor chunk system

The following classes are the most important in the anchor chunk system:

- **InputText**
  The class *InputText* defines a parallel text as a sequence of sentence pairs. The class contains the method *processInputText*. This method reads aligned sentence pairs from the sentence-aligned input texts and processes each sentence pair. The method also reads the text files containing the English and Dutch additional linguistic annotations (see section 2.5).

- **SentencePair**
  The class *SentencePair* defines a bilingual sentence pair. The sentence pairs are read from a sentence-aligned input text file. The SentencePair class is the most crucial class in the process as for each sentence pair, a lexical link matrix is built and the anchor chunks are linked.

- **Sentence**
  The class *Sentence* defines a monolingual sentence. Both the source and target sentence of the SentencePair class are instances of the Sentence class. A sentence consists of a list of words (tokens) in a given language. Each sentence is chunked (see section 2.5.3). This chunk information is stored on the one hand as a set of chunk boundaries. On the other hand, each sentence also contains a list of chunks.

- **Word**
  The class *Word* defines a word. The class contains monolingual information: word form (token), lemma and PoS information. Each word is looked up in the bilingual lexicon and all possible translations together with the translation probabilities are retrieved.

- **WordLink**
  The class *WordLink* defines a link between a source and target word. A word link is defined by the position of the source word in the source sentence and the position of the target word in the target sentence and the types of link (e.g. lexical link, part-of-speech link).

- **Chunk**
  The class *Chunk* defines a chunk as a sublist of a Sentence. A chunk consists of a sequence of Words.

- **Lexicon**
  The class *Lexicon* contains a method to read a text file containing the bilingual lexicon. The bilingual lexicon must contain one line for each dictionary entry. Each dictionary entry consists of an English word, the corresponding Dutch word and optionally a probability value. The different fields are delimited by the #-symbol.

Figure C.1 vizualises the most important classes of the Anchor Chunk system in UML. UML stands for Unified Modeling Language. It is a widely used graphical representation scheme for modelling object-oriented systems.

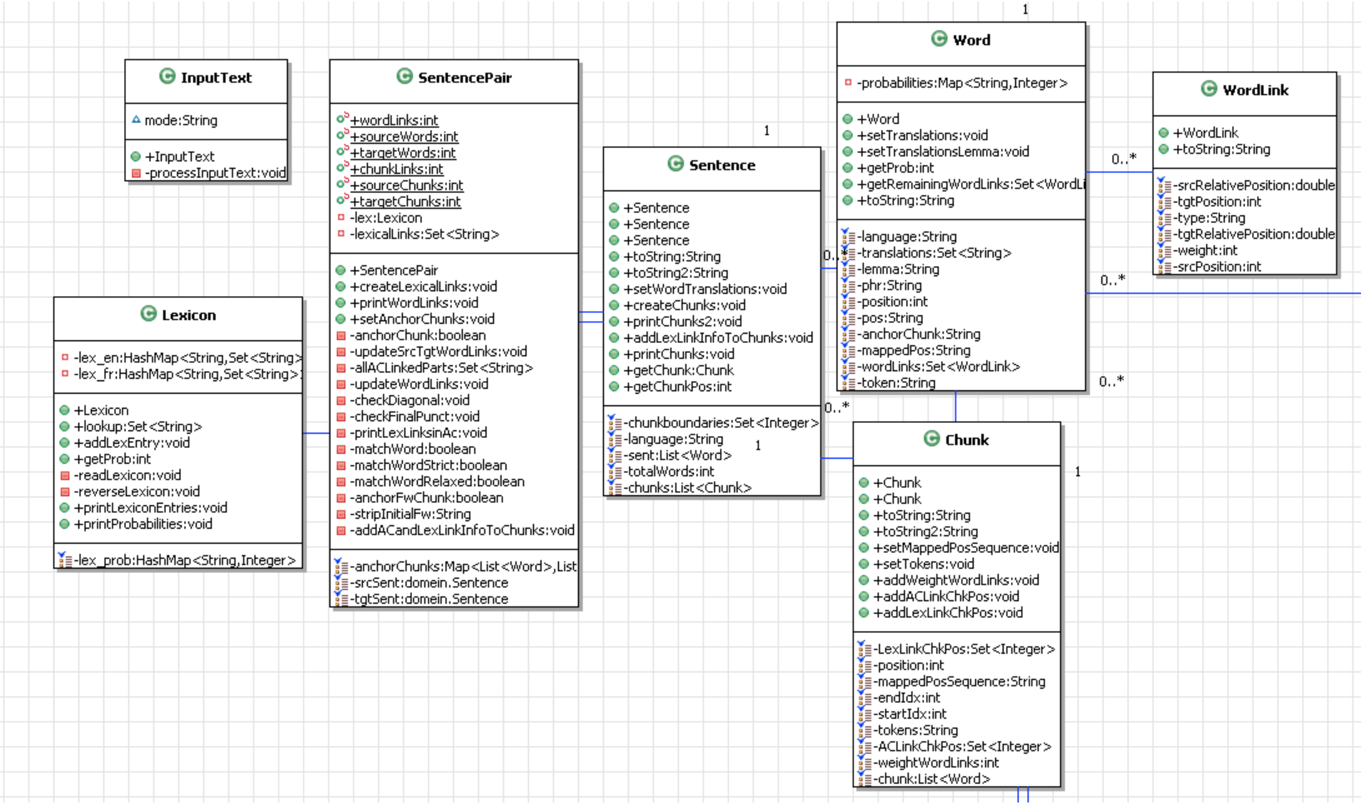


Figure C.1: Graphical representation of the most important classes of the anchor chunk system.

C.2 Bootstrapping system

The bootstrapping system builds further on the anchor chunk system. For the bootstrapping system, two versions of the class *InputText* were created: *InputTextTrain* and *InputTextValidate*. The bootstrapping process is a cyclic process that alternates between extracting candidate translation rules (extraction step) – which is implemented in the class *InputTextTrain* and scoring and filtering the extracted candidate translation rules (validation step) – which is implemented in the class *InputTextValidate*.

The following classes are the most important in the bootstrapping system:

- **TranslationRule**
  The class *TranslationRule* defines a translation rule. A translation rule is either defined by a source and a target pattern (for *abstract* translation rules) or by its source and target tokens (for *lexicalized* translation rules). The class *TranslationRule* contains a number of attributes to calculate a number of statistics (e.g. Log-Likelihood ratio, number of times a rule was extracted or applied, and the like).

- **FiredTranslationRule**
  The class *FiredTranslationRule* contains an instance of a translation rule together with pointers to the source and target chunks that are linked. Instances of the class *FiredTranslationRule* are used in the validation method that uses a manually created reference corpus and in the test environment (see section C.3).

- **TranslationRules**
  The class *TranslationRules* contains a list of instances of *TranslationRule*. The class *TranslationRules* is used to store all extracted rules in the extraction step and all validated rules in the validation step.

- **Bootstrapping**
  The class *Bootstrapping* only contains all methods that are used in the bootstrapping process: extractRules, applyRules (which calls applyLexicalRule and applyPosRule) and validateTranslationRules.

- **GizaAlignments** and **GizaWordAlignments**
  In the bootstrapping system, the classes *GizaAlignments* and *GizaWordAlignments* are used instead of the *Lexicon* class. The GIZA++ word alignments are stored per sentence pair as a set of source-target word indices.

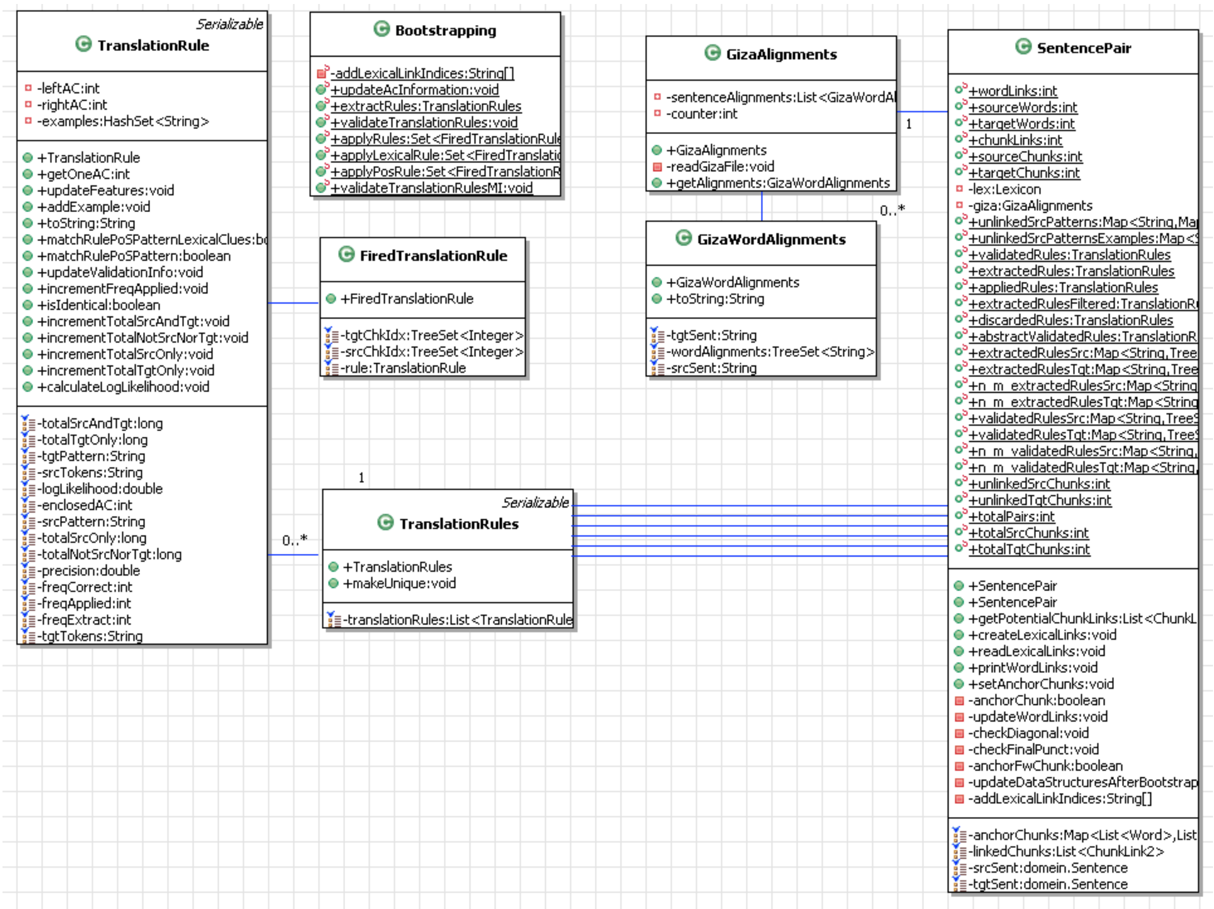

Serializable
TranslationRule
-leftAC:int
-rightAC:int
-examples:HashSet<String>
+TranslationRule
+getOneAC:int
+updateFeatures:void
+addExample:void
+toString:String
+matchRulePoSPatternLexicalClues:bo
+matchRulePoSPattern:boolean
+updateValidationInfo:void
+incrementFreqApplied:void
+isIdentical:boolean
+incrementTotalSrcAndTgt:void
+incrementTotalNotSrcNorTgt:void
+incrementTotalSrcOnly:void
+incrementTotalTgtOnly:void
+calculateLogLikelihood:void
-totalSrcAndTgt:long
-totalTgtOnly:long
-tgtPattern:String
-srcTokens:String
-logLikelihood:double
-enclosedAC:int
-srcPattern:String
-totalSrcOnly:long
-totalNotSrcNorTgt:long
-precision:double
-freqCorrect:int
-freqApplied:int
-freqExtract:int
-tgtTokens:String
Bootstrapping
-addLexicalLinkIndices:String[]
+updateAcInformation:void
+extractRules:TranslationRules
+validateTranslationRules:void
+applyRules:Set<FiredTranslationRule
+applyLexicalRule:Set<FiredTranslatio
+applyPosRule:Set<FiredTranslationR
+validateTranslationRulesMI:void
FiredTranslationRule
+FiredTranslationRule
-tgtChkIdx:TreeSet<Integer>
-srcChkIdx:TreeSet<Integer>
-rule:TranslationRule
1
0..*
Serializable
TranslationRules
+TranslationRules
+makeUnique:void
-translationRules:List<TranslationRule
GizaAlignments
-sentenceAlignments:List<GizaWordAl
-counter:int
+GizaAlignments
-readGizaFile:void
+getAlignments:GizaWordAlignments
1
0..*
GizaWordAlignments
+GizaWordAlignments
+toString:String
-tgtSent:String
-wordAlignments:TreeSet<String>
-srcSent:String
SentencePair
+wordLinks:int
+sourceWords:int
+targetWords:int
+chunkLinks:int
+sourceChunks:int
+targetChunks:int
-lex:Lexicon
-giza:GizaAlignments
+unlinkedSrcPatterns:Map<String,Ma
+unlinkedSrcPatternsExamples:Map<S
+validatedRules:TranslationRules
+extractedRules:TranslationRules
+appliedRules:TranslationRules
+extractedRulesFiltered:TranslationRu
+discardedRules:TranslationRules
+abstractValidatedRules:TranslationR
+extractedRulesSrc:Map<String,Tree
+extractedRulesTgt:Map<String,Tree
+n_m_extractedRulesSrc:Map<String
+n_m_extractedRulesTgt:Map<String
+validatedRulesSrc:Map<String,TreeS
+validatedRulesTgt:Map<String,TreeS
+n_m_validatedRulesSrc:Map<String,
+n_m_validatedRulesTgt:Map<String,
+unlinkedSrcChunks:int
+unlinkedTgtChunks:int
+totalPairs:int
+totalSrcChunks:int
+totalTgtChunks:int
+SentencePair
+SentencePair
+getPotentialChunkLinks:List<ChunkL
+createLexicalLinks:void
+readLexicalLinks:void
+printWordLinks:void
+setAnchorChunks:void
-anchorChunk:boolean
-updateWordLinks:void
-checkDiagonal:void
-checkFinalPunct:void
-anchorFwChunk:boolean
-updateDataStructuresAfterBootstrap
-addLexicalLinkIndices:String[]
-anchorChunks:Map<List<Word>,List
-linkedChunks:List<ChunkLink2>
-srcSent:domein.Sentence
-tgtSent:domein.Sentence

Figure C.2: Graphical representation of the most important classes of the bootstrapping system.

C.3 Test system

In the test environment, the validate rules are applied on all the sentence pairs after anchor chunk alignment. The resulting alignments are compared with the manually created reference alignments to calculate Weighted F-measure, Alignment Error Rate and F-measure at chunk level.

List of Figures

List of Tables

Bibliography

Abney, S.: 1991, Parsing by chunks, *in* R. Berwick, S. Abney and C. Tenny (eds), *Principle-Based Parsing*, Kluwer Academic Publishers, pp. 257–278.

Ahmad, K., Gillam, L. and Tostevin, L.: 1999, University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER), *Proceedings of the Eight Text REtrieval Conference (TREC-8)*, Maryland, pp. 717–724.

Ahrenberg, L., Andersson, M. and Merkel, M.: 2002, A system for incremental and interactive word linking, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, pp. 485–490.

Asadi, P. and Séguinot, C.: 2005, Shortcuts, strategies and general patterns in a process study of nine professionals, *Meta* **50**(2), 522–547.

Baayen, R., Piepenbrock, R. and van Rijn, H.: 1993, The CELEX lexical database on CD-ROM.

Bakker, M., Koster, C. and Van Leuven-Zwart, K.: 2008, Shifts, *in* M. Baker and G. Saldanha (eds), *Routledge Encyclopedia of Translation Studies*, Routledge, London, New York, pp. 269–274.

Barkhudarov, L.: 1993, The problem of the unit of translation, *in* P. Zlateva (ed.), *Translation as social action: Russian and Bulgarian perspectives*, Routledge, London, pp. 39–46.

Bernardini, S.: 2001, Think-aloud protocols in translation research: achievements, limits, future prospects, *Target* **13**(2), 241–263.

Bourigault, D.: 1992, Surface grammatical analysis for the extraction of terminological noun phrases, *Proceedings of the 15th conference on Computational linguistics (Coling)*, Nantes, France, pp. 977–981.

Bowker, L.: 2008, Terminology, *in* M. Baker and G. Saldanha (eds), *Routledge Encyclopedia of Translation Studies*, Routledge, London, New York, pp. 286–290.

Bowker, L. and Barlow, M.: 2004, Bilingual concordancers and translation memories: A comparative evaluation, *Proceedings of the second International Workshop on Language Resources for Translation Work, Research and Training (Coling)*, Geneva, Switzerland, pp. 52–61.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A. and Mercer, R. L.: 1993, The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics* **19**(2), 263–311.

Brown, P. F., Lai, J. C. and Mercer, R. L.: 1991, Aligning sentences in parallel corpora, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, pp. 169–176.

Buchholz, S. and Daelemans, W.: 2001, Complex answers: a case study using a WWW question answering system., *Natural Language Engineering* **7**, 301–323.

Buhmann, J., Martens, J.-P., Macken, L. and Van Coile, B.: 2002, Intonation Modelling for the Synthesis of Structured Documents, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado (USA), pp. 2089–2092.

Cabré Castellví, T.: 2003, Theories of terminology. Their description, prescription and explanation, *Terminology* **9**(2), 163–199.

Cabré Castellví, T., Bagot, R. E. and Palatresi, J. V.: 2001, Automatic term detection. A review of current systems., *in* D. Bourigault, C. Jacquemin and M.-C. L'Homme (eds), *Recent advances in computational terminology*, John Benjamins, Amsterdam, pp. 149–166.

Carl, M.: 2009, Grounding Translation Tools in Translator's Activity Data, *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.

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

Catford, J, C.: 1965, *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, Oxford University Press, Oxford.

Cyrus, L.: 2006, Building a resource for studying translation shifts, *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, Genua, Italy.

Daelemans, W. and van den Bosch, A.: 2005, Application to shallow parsing, *in* W. Daelemans and A. Van den Bosch (eds), *Memory-based language processing*, Cambridge University Press, Cambridge, United Kingdom, pp. 85–103.

Daille, B.: 1996, Study and implementation of combined techniques for automatic extraction of terminology, *in* J. L. Klavans and P. Resnik (eds), *The balancing act: combining symbolic and statistical aproaches to language*, MIT Press, Massachusetts.

Danielsson, P. and Ridings, D.: 1997, Practical presentation of a "vanilla" aligner, *Proceedings of the TELRI Workshop on Alignment and Exploitation of Texts*, Ljubljana.

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

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

Davis, P. C., Xie, Z. and Small, K.: 2007, All Links are not the Same: Evaluating Word Alignments for Statistical Machine Translation, *in* B. Maegaard (ed.), *Machine Translation Summit XI*, European Associaton for Machine Translation, Copenhagen, Denmark, pp. 119–126.

De Groot, F.: 2006, Personal Communication, E-mail 18-05-2006.

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

Dias, G. and Kaalep, H.: 2003, Automatic extraction of multiword units for Estonian: Phrasal verbs, *Languages in Development* **41**, 81–91.

Dunning, T.: 1993, Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics* **19**(1), 61–74.

Ericsson, K. A. and Simon, H. A.: 1993/1984, *Protocol Analysis. Verbal reports as data*, MIT Press, Cambridge, MA.

Frantzi, K. and Ananiadou, S.: 1999, The C-value / NC-value domain independent method for multi-word term extraction, *Journal of Natural Language Processing* **6**(3), 145–179.

Fulford, H.: 2001, Exploring terms and their linguistic environment. A domain-independent approach to automated term extraction, *Terminology* **7**(2), 259–279.

Gale, W. A. and Church, K. W.: 1991, A program for aligning sentences in bilingual corpora, *Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp. 177–184.

Gale, W. A. and Church, K. W.: 1993, A program for aligning sentences in bilingual corpora, *Computational Linguistics* **19**(1), 75–102.

Gamper, J. and Stock, O.: 1999, Corpus-based terminology, *Terminology* **5**(2), 147–159.

Gaussier, E.: 1998, Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL)*, Université de Montréal, Montreal, Quebec, Canada, pp. 444–450.

Goetschalckx, J., Cucchiarini, C. and Van Hoorde, J.: 2001, Machine Translation for Dutch: the NL-Translex Project.

Gotti, F., Langlais, P., Macklovitch, E., Bourigault, D., Robichaud, B. and Coulombe, C.: 2005, 3GTM: a third-generation translation memory, *Proceedings of the third Computational Linguistics in the North-East (CLiNE) Workshop*, Gatineau, Québec.

Hatim, B. and Munday, J.: 2004, *Translation. An advanced resource book*, Routledge, Oxon, New York.

Isabelle, P., Dymetman, M., Foster, G., Jutras, J.-M., Macklovitch, E., Perrault, F., Ren, X. and Simard, M.: 1993, Translation Analysis and Translation Automation, *TMI-93: The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, pp. 201–217.

Itagaki, M., Aikawa, T. and He, X.: 2007, Automatic Validation of Terminology Consistency with Statistical Method, *in* B. Maegaard (ed.), *Machine Translation Summit XI*, European Associaton for Machine Translation, Copenhagen, Denmark, pp. 269–274.

Jääskeläinen, R.: 2002, Think-aloud protocols into translation: an annotated bibliography, *Target* **14**(1), 107–136.

Jakobsen, A. L.: 2005, Investigating expert translators' processing knowledge, *in* H. V. Dam, J. Engberg and H. Gerzymiscg-Arbogast (eds), *Knowledge systems and translation*, Walter de Gruyter, Berlin, pp. 173–189.

Justeson, K. and Katz, S.: 1995, Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering* **1**(1), 9–27.

Kageura, K. and Umino, B.: 1996, Methods of automatic term recognition. A review, *Terminology* **3**(2), 259–289.

Kay, M. and Röscheisen, M.: 1993, Text-Translation Alignment, *Computational Linguistics* **19**(1), 121–142.

Ker, S. J. and Chang, J. S.: 1997, A Class-Based Approach to Word Alignment, *Computational Linguistics* **23**(2), 313–343.

Koehn, P.: 2005, Europarl: a parallel corpus for statistical machine translation, *Tenth Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Koehn, P.: 2009, *Statistical Machine Translation*, Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: 2007, Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180.

Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y. and Cofino, T.: 2004, Glossary extraction and utilization in the information search and delivery system for IBM Technical Support, *IBM System Journal* **43**(3).

Kupiec, J.: 1993, An algorithm for finding noun phrase correspondences in bilingual corpora, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, Columbus, Ohia, United States.

Kurokawa, D., Goutte, C. and Isabelle, P.: 2009, Automatic Detection of Translated Text and its Impact on Machine Translation, *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, Ottawa, Ontario, Canada.

Lagoudaki, E.: 2006, Translation Memory systems: Enlightening users' perspective. Key findings of the TM Survey 2006, *Technical report*, Imperial College London.

Lagoudaki, E.: 2008, The value of machine translation for the professional translator, *Proceedings of the eighth Conference of the Association for Machine Translation in the Americas (AMTA-2008)*, Waikiki, Hawaii, pp. 262–269.

Lambert, P., De Gispert, A., Banchs, R. and Mariño, J. B.: 2005, Guidelines for Word Alignment Evaluation and Manual Alignment, *Language Resources and Evaluation* **39**, 267–285.

Lee, D. Y. W.: 2001, Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle, *Language Learning & Technology* **5**(3), 37–72.

Lefever, E., Macken, L. and Hoste, V.: 2009, Language-independent bilingual terminology extraction from a multilingual parallel corpus, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Athens, Greece, pp. 496–504.

Liu, Y. and Lin, S.: 2005, Log-linear models for word alignment, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, pp. 459–466.

Lommel, A.: 2004, LISA 2004 Translation Memory Survey.

Lörscher, W.: 1992, Investigating the Translation Process, *Meta* **37**(3), 426–439.

Lörscher, W.: 2005, The translation process: methods and problems of its investigation, *Meta* **50**(2), 597–608.

Macken, L.: 2007, Analysis of translational correspondence in view of subsentential alignment, *in* F. Van Eynde, V. Vandeghinste and I. Schuurman (eds), *METIS-II Workshop on New Approaches to Machine Translation*, Centre for Computational Linguistics, University of Leuven, Leuven, Belgium, pp. 97–105.

Macken, L.: 2009, In search of the recurrent units of translation, *in* W. Daelemans and V. Hoste (eds), *Evaluation of Translation Technology. LANS 8/2009*, Evaluation of Translation Technology. LANS 8/2009, Brussels, Belgium, pp. 195–212.

Macken, L.: 2010, An annotation scheme and Gold Standard for Dutch-English word alignment, *Proceedings of the seventh Language Resources and Evaluation Conference (LREC)*, Valletta, Malta.

Macken, L. and Daelemans, W.: 2009, Aligning linguistically motivated phrases, *in* S. Verberne, H. van Halteren and P.-A. Coppen (eds), *Computational Linguistics in the Netherlands 2007: Selected papers from the eighteenth CLIN meeting*, LOT Occasional Series 11, Utrecht, The Netherlands, pp. 37–52.

Macken, L. and Daelemans, W.: 2010, A Chunk-Driven Bootstrapping Approach to Extracting Translation Patterns, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (Iasi, Romania)*, Vol. 6009 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 394–405.

Macken, L., Lefever, E. and Hoste, V.: 2008, Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*, Manchester, United Kingdom, pp. 529–536.

Macken, L., Trushkina, J. and Rura, L.: 2007, Dutch Parallel Corpus: MT corpus and translator's aid, *in* B. Maegaard (ed.), *Machine Translation Summit XI*, European Associaton for Machine Translation, Copenhagen, Denmark, pp. 313–320.

Magerman, D. M. and Marcus, M. P.: 1990, Parsing a Natural Language Using Mutual Information Statistics, *Proceedings of the eighth National Conference on Artificial Intelligence (AAAI-90)*, pp. 984–989.

Manning, C. D. and Schütze, H.: 2003, *Foundations of Statistical Natural Language Processing*, Massachusetts Institute of Technology.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A.: 1993, Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics* **19**(2), 313–330.

McEnery, T., Richard, X. and Tono, Y.: 2006, *Corpus-based Language Studies. An advanced resource book*, Routledge Applied Linguistics, Routledge, London.

Melamed, D. I.: 1997, A Portable Algorithm for Mapping Bitext Correspondence, *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)*, Madrid, Spain, pp. 305–312.

Melamed, D. I.: 1998, Empirical methods for MT lexicon development, *in* D. Farwell, L. Gerber and E. Hovy (eds), *Proceedings of the third Conference of the Association for Machine Translation in the Americas (AMTA)*, Springer-Verlag, Langhorne PA, USA.

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

Melamed, D. I.: 2001a, Annotation style guide for the Blinker Project, *in* D. I. Melamed (ed.), *Empirical methods for exploiting parallel texts*, MIT Press, Cambridge, Massachusetts, pp. 169–182.

Melamed, D. I.: 2001b, *Empirical methods for exploiting parallel texts*, MIT Press, Cambridge, Massachusetts.

Melamed, D. I.: 2001c, Manual annotation of translational equivalence, *in* D. I. Melamed (ed.), *Empirical methods for exploiting parallel texts*, MIT Press, Cambridge, Massachusetts, pp. 65–77.

Merkel, M.: 1999a, Annotation Style Guide for the PLUG Link Annotator, *Technical report*, Department of Computer and Information Science, Linköping University.

Merkel, M.: 1999b, *Understanding and enhancing translation by parallel text processing*, PhD thesis, Linköping university.

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

Moore, R. C.: 2002, Fast and accurate sentence alignment of bilingual corpora, *Proceedings of the fifth Conference of the Association for Machine Translation in the Americas (AMTA)*, Machine Translation: from research to real users, Tiburon, California, pp. 135–244.

Moore, R. C.: 2005, Association-Based Bilingual Word Alignment, *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Association for Computational Linguistics, Ann Arbor, Michigan, pp. 1–8.

Newmark, P.: 1988, *A Textbook of Translation*, Prentice Hall, New York/London.

Och, F. J. and Ney, H.: 2000, A comparison of alignment models for statistical machine translation, *Proceedings of the 18th International Conference on Computational Linguistics (Coling)*, Saarbrücken, Germany.

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

Odijk, J., Martens, J.-P., Van Eynde, F., Daelemans, W., Kenyon-Jackson, D., Vossen, P., van Hesse, A., Boves, L. and Beeken, J.: 2004, *Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie. STEVIN. Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands*, Nederlandse Taalunie, The Hague.

Olohan, M.: 2004, *Introducing Corpora in Translation Studies*, Routledge, London/New York.

Park, Y., Byrd, R. J. and Boguraev, B. K.: 2002, Automatic glossary extraction: beyond terminology identification, *Proceedings of the 19th international conference on Computational linguistics (Coling)*, Taipei, Taiwan.

Planas, E.: 2005, SIMILIS. Second-generation translation memory software, *Proceedings of the 27th International Conference on Translating and the Computer (TC27)*, ASLIB, London, United Kingdom.

Planas, E. and Furuse, O.: 2003, Formalizing translation memory, *in* M. Carl and A. Way (eds), *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, pp. 157–188.

Quin, D.: 1997, Terminology for Machine Translation: a Study, *Machine Translation Review* **6**, 9–21.

Rayson, P. and Garside, R.: 2000, Comparing corpora using frequency profiling, *Proceedings of the workshop on Comparing corpora, 38th annual meeting of the Association for Computational Linguistics*, Hong Kong, China, pp. 1–6.

Schmid, H.: 1994, Probabilistic part-of-speech tagging using decision trees, *Proceedings of the International Conference on new methods in Language Processing*, Manchester, UK.

Sclano, F. and Velardi, P.: 2007, TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities, *Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007)*, Sophia Antinopolis.

Simard, M., Foster, G., Hannan, M.-L., Macklovitch, E. and Plamondon, P.: 2000, Bilingual text alignment: where do we draw the line?, *in* S. Botley, A. McEnery and A. Wilson (eds), *Multilingual corpora in teaching and research*, Rodopi, Amsterdam, pp. 38–64.

Simard, M. and Langlais, P.: 2001, Sub-sentential exploitation of translation memories, *MT Summit VIII*, Santiago De Compostela, Spain.

Simard, M. and Macklovitch, E.: 2005, Studying the human translation process through the TransSearch log-files, *Proceedings of the AAAI Symposium on Knowledge Collection from volunteer contributors*, Stanford, California, USA.

Smadja, F., McKeown, K. and Hatzivassiloglou, V.: 1996, Translating collocations for bilingual lexicons: A statistical approach, *Computational Linguistics* **22**(1), 1–38.

Spakov, O. and Räihä, K.-J.: 2008, KiEV: A tool for visualization of reading and writing processes while translating a text, *Eye Tracking Research & Applications Symposium (ETRA08)*, Savannah, GA, pp. 107–110.

Sparck Jones, K.: 1979, Experiments in relevance weighting of search terms, *Information Processing and Management* **15**, 133–144.

Thunes, M.: 1998, Classifying translational correspondences, *in* S. Johansson and S. Oksefjell (eds), *Corpora and cross-linguistic research: theory, method and case studies*, Rodopi, Amsterdam, Atlanta, pp. 25–50.

Tiedemann, J.: 2001, Can bilingual word alignment improve monolingual phrasal term extraction?, *Terminology* **7**(2), 199–215.

Tjong Kim Sang, E. F. and Buchholz, S.: 2000, Introduction to the CoNLL-2000 Shared Task: Chunking, *CoNLL-2000 and LLL-2000*, Lisbon, Portugal, pp. 127–132.

van den Bosch, A., Busser, B., Daelemans, W. and Canisius, S.: 2007, An efficient memory-based morphosyntactic tagger and parser for Dutch, *Computational Linguistics in the Netherlands 2006*, Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting, Leuven, Belgium, pp. 191–206.

van den Bosch, A., Schuurman, I. and Vandeghinste, V.: 2006, Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development, *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, Genua, Italy.

Van Eynde, F., Zavrel, J. and Daelemans, W.: 2000, Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus, *Proceedings of the second International Conference on Language Resources and Evaluation (LREC)*, Vol. III, Athens, Greece, pp. 1427–1433.

Van Leuven-Zwart, K.: 1989, Translation and original: Similarities and dissimilarities I, *Target* **1**(2), 151–181.

Van Leuven-Zwart, K.: 1990, Translation and original: Similarities and dissimilarities II, *Target* **2**(1), 69–95.

Vandeghinste, V.: 2008, *A Hybrid Modular Machine Translation System. LoRe-MT: Low Resources Machine Translation*, PhD thesis, Centre for Computational Linguistcs, K.U. Leuven.

Véronis, J.: 1998, Arcade. Tagging guidelines for word alignment. Version 1.0, *Technical report*.

Véronis, J.: 2000, Evaluation of parallel text alignment systems: the ARCADE project, *in* J. Véronis (ed.), *Parallel text processing: alignment and use of translation corpora*, Kluwer Academic Publishers, Dordrecht, pp. 369–388.

Vinay, J.-P. and Darbelnet, J.: 1958, *Stylistique Comparée du Franais et de lAnglais: Méthode de Traduction*, Didier, Paris.

Wright, S. E.: 1997, Term selection: the initial phase of terminology management, *in* S. E. Wright and G. Budin (eds), *Handbook of terminology management*, John Bejamins, Amsterdam, pp. 13–23.

Zhang, Z., Iria, J., Brewster, C. and Ciravegna, F.: 2008, A comparative evaluation of term recognition algorithms, *Proceedings of the sixth international conference of Language Resources and Evluation (LREC)*, Marrakech, Morocco.

Index