Linköping Studies in Science and Technology

Dissertation No. 607

Understanding and enhancing translation by parallel text processing

by

Magnus Merkel

Department of Computer and Information Science
Linköpings universitet
SE-581 83 Linköping, Sweden

Linköping 1999

Printed by UniTryck, Linköping 1999

Abstract

In recent years the fields of translation studies, natural language processing and corpus linguistics have come to share one object of study, namely parallel text corpora, and more specifically translation corpora. In this thesis it is shown how all three fields can benefit from each other, and, in particular, that a prerequisite for making better translations (whether by humans or with the aid of computer-assisted tools) is to understand features and relationships that exist in a translation corpus. The Linköping Translation Corpus (LTC) is the empirical foundation for this work. LTC is comprised of translations from three different domains and translated with different degrees of computer support. Results in the form of tools, measures and analyses of translations in LTC are presented.

In the translation industry, the use of translation memories, which are based on the concept of reusability, has been increasing steadily in recent years. In an empirical study, the notion of reusability in technical translation is investigated as well as translators' attitudes towards translation tools.

A toolbox for creating and analysing parallel corpora is also presented. The tools are then used for uncovering relationships between the originals and their corresponding translations. The Linköping Word Aligner (LWA) is a portable tool for linking words and expressions between a source and target text. LWA is evaluated with the aid of reference data compiled before the system evaluation. The reference data are created and evaluated automatically with the help of an annotation tool, called the PLUG Link Annotator.

Finally, a model for describing correspondences between a source text and a target text is introduced. The model uncovers voluntary shifts concerning structure and content. The correspondence model is then applied to the LTC.

Acknowledgements

There are a whole lot of people out there who have had their part in this work:

First and foremost, I wish to express my thanks to my adviser, Lars Ahrenberg, for being a fantastic discussion partner, showing divine patience and contributing to numerous improvements on many versions and drafts of this thesis.

Many people have been involved in developing the software, but there are two people in particular that I am indebted too: Mikael Andersson and Bernt Nilsson. Mikael, thanks for your help with LWA and Frasse-2. Bernt, a big thank you for the time and effort put into Frasse-1 and the first version of the sentence aligner.

I thank all past and present members of the Natural Language Processing Laboratory for making this place such a creative and inspiring environment: Lars Ahrenberg, Nils Dahlbäck, Robert Eklund, Annika Flycht-Eriksson, Richard Hirsch, Peter Ingels, Arne Jönsson, Anders Lindström, Bertil Lyberg, Pernilla Qvarfordt, Mustapha Skhiri, Lena Strömbäck, Stefan Svenberg, Åke Thurée and Mats Wirén. Among these people, the initial members of NLPLAB cannot be thanked enough: Lars Ahrenberg, Nils Dahlbäck and Mats Wirén for their encouragement during the early years of NLPLAB and especially Arne Jönsson for many relevant and many not so relevant discussions on topics ranging from AI to more mundane, but nevertheless never boring, subjects. Arne also deserves a special thank you for his last minute comments to this thesis.

I am also indebted to people outside the department who have contributed in various ways. Thanks to Liesbeth van Bijsterveld and Gunnie Jacobsson at Microsoft for providing me with electronic versions of the manuals. Thanks to Peter Bursell who managed to do the same at IBM. Thanks to Martin Gellerstam at Språkbanken who helped me with the fiction material, and to Bengt Altenberg in Lund who provided information about the ESP Corpus.

I would also like to thank all the people in the PLUG project from Uppsala and Gothenburg for valuable discussions. Thanks to Anna Sågvall Hein, Jörg Tiedemann, Daniel Ridings and Katarina Mühlenbock.

A big thank you also goes to Ivan Rankin who improved the English in this thesis considerably.

Thanks to all students and project workers who have been around over the years, especially Sara Nordberg, who implemented the first version of the discrepancy tool, as well as Mathias Holm and Markus Olsson who designed the graphical version of the DAVE toolbox presented in the thesis.

I am also grateful to all the people in the translation business who filled in the questionnaires and who took time to answer questions. Special thanks go to Svante Skoglund, Charlotte Lotoft, Åsa Karlsson and Henrik Lundström.

I would also like to thank Lillemor Wallgren, Marie Eklund, Lisbeth Linge and Lena Wigh for their administrative support and for reminding me of things I always seem to forget.

The technical staff at the department also deserves recognition. Thanks to Leif Finnmo, Arne Fäldt, Peter J. Nilsson, Bernt Nilsson, Rolf Nilsson and Göran Sedvall for the excellent technical support.

Thanks also to everybody else at the department who directly or indirectly supported this work, especially Patrick Doherty, Inger Emanuelsson, Anders Haraldsson, Sture Hägglund, Erik Sandewall and Jalal Maleki.

The research presented in this thesis was supported by the Swedish National Board for Industrial and Technical Development (NUTEK) and the Swedish Council for Research in the Humanities and Social Sciences (HSFR).

Finally, I thank all friends and relatives and especially my fantastic children. Thank you Annabel and Axel for just being who you are.

Linköping October 1999

Contents

1 Introduction

In recent years the fields of translation studies, natural language processing and corpus linguistics have come to share one object of study, namely *parallel text corpora*, and more specifically *translation corpora* (see Figure 1 below). Translation scholars aim at uncovering relationships and characteristics in translations. Computational linguists have an interest in developing better algorithms and tools that will improve machine translation systems and aid translators in their work. Another objective within natural language processing is to extract data from corpora that can be included in applications, for example, machine-readable lexicons and terminological data. Corpus linguists are interested in methods for creating and exploring text corpora for applications such as contrastive linguistics and lexicography.

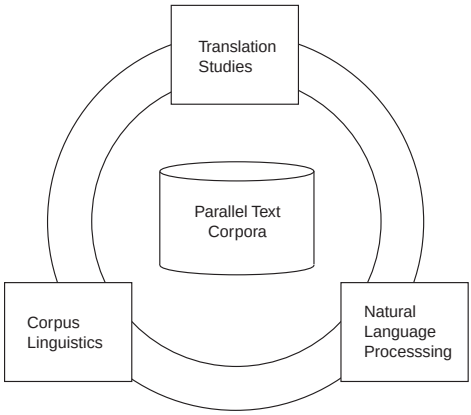


Figure 1. Parallel text corpora as an empirical foundation for translation studies, natural language processing and corpus linguistics

In this work I wish to stress the relationships between the different fields of inquiry. A key idea is that one can only build high-quality *translation systems* and *translation support tools* if the system designers know what characterises high-quality translation. To be able to characterise what high quality translations are, theories and methods from the field of translation studies should be used. However, the empirical foundation is to be found in existing translations, as Isabelle et al. have pointed out:

> … existing translations contain more solutions to more translation problems than any other available resource. (Isabelle et al. 1993, p. 205).

One can take Isabelle's claim one step further and say that existing translations contain more solutions to *problems for translation systems* than any other available resource. In other words, it is by *understanding* translation, as it is manifested in existing translations, that language engineers can *enhance* translation tools and automatic translation systems. The other side of the coin is that tools and systems that are designed in the NLP field can also be used by translation scholars to test hypotheses and for developing translation theory. For example, from the NLP perspective, a certain tool could have been built as an evaluation tool for finding mistakes in industrialised translation, but the same tool can also be of help for the translation researcher to uncover relationships between a source text and a target text. Although the objectives may differ among the research fields, tools originally designed for one task can often be applied for different purposes. Acquiring a better knowledge of translation from empirical sources not only entails improved translation software, but other areas such as translation training, lexicography and contrastive linguistics could also reap the translation harvest.

The idea to find solutions to translation problems by using previous translations is already manifested in a new generation of translation software using *translation memories*. However, successful applications of translation memories rest to a large extent on the concept of *reusability*, i.e., that translations can be recycled and used over and over again. In this work, *reusability* as it concerns translation is investigated from the perspective from the translator. Another vital theme in this thesis is the notion of *correspondence*, which will be adressed from different angles as it is fundamental when relationships between a source text and its target text are investigated on all levels.

1.1 Goal

The overall objective of this thesis is to show how translations can be produced more efficently and with improved quality if parallell texts and parallel text processing tools are utilized.

In order to achieve this objective, certain smaller steps have been taken:

1. Collect translations and prepare them as sentence aligned parallel texts (bitexts).

2. Develop tools for the analysis of monolingual texts (recurrence data and collocation retrieval).

3. Investigate translators' attitudes towards translation tools and the notion of reusability, which is a fundamental feature of translation memory-based translation.

4. Investigate, and describe, the characteristics of different translations (human translations, semi-automatic translations and automatic translations) and different text types (technical manuals and fiction).

5. Develop and evaluate tools for the analysis of bilingual texts (discrepancy data, bilingual concordancing and bilingual word alignment).

6. Develop a model in which the structural and semantic correspondence can be described.

1.2 Contributions

The main contributions of this thesis are the following:

- Integration of techniques and methods from translation studies, corpus linguistics and natural language processing in order to improve the understanding of translation and, consequently, create a platform for building better translation tools and translation systems.

- Results on how translation relationships differ with text types and translation methods.

- Linköping Word Aligner, a portable high-performance word alignment system.

- Development and application of a correspondence model for characterising structural and semantic change in translations.

- An investigation into translators' attitudes towards translation tools in general, and towards reusability in particular. The results provide a possible explanation to why translation memory-based translation has not lived up to promises in efficiency and performance.

Other contributions include:

- A sentence-aligned English-Swedish translation corpus: The Linköping Translation Corpus.

- A parallel text processing solution capable of handling the relationship between Scandinavian languages such as Swedish and other western-European languages.

- A set of resource building tools, diagnostic tools, extraction tools, and evaluation tools used to discover relevant features of the translation corpus. The tools require little or no language-specific resources and are therefore extremely portable.

- A set of measures related to translation relationships concerning recurrence, consistency and co-occurrence.

- A method and software for evaluating word-alignment systems.

1.2.1 Overview of thesis

The outline of the thesis is as follows:

Chapter 2: A background to the relevant areas within translation studies, corpus linguistics and computational linguistics is presented, which will set the ground for a joint view of translation corpora as a common source of study for translation studies, contrastive linguists, translators and computational linguists.

Chapter 3: The notion of "reusability", a cornerstone for translation-memory-based translation, is examined via an empirical study of translators' attitudes towards consistency. The chapter sets the ground for how translation corpora (and translation memories) can be used by translators in practice. The possible pitfalls and advantages of such an approach are presented.

Chapter 4: The Linköping Translation Corpus (LTC) and analyses of source and target texts independently are presented.

Chapter 5: The Dave toolbox is presented. Dave consists of a set of tools related to translation with knowledge-lite, string-based processing which helps to diagnose, monolingual texts and parallel texts, build resources, extract data and evaluate translations. Its usability for translators, contrastive linguists, language engineers and translation scholars is described.

Chapter 6: Analyses of the Linköping Translation Corpus with parallel text processing. Data on sentence mappings, consistency and variation on the sentence level, lexical co-occurrences are presented as well as the implications of such analyses.

Chapter 7: The Linköping Word Aligner (LWA) is presented along with assumptions underlying its design, the architecture and the knowledge-lite modules. First a description of LWA's application on English/Swedish translations is given. Then the portability of LWA is illustrated with a second version for French/English.

Chapter 8: Different approaches to the evaluation of word alignment systems are discussed. A system for interactive annotation of reference data used in word alignment evaluations is presented (the PLUG Link Annotator).

Chapter 9: Three separate evaluations of LWA are presented as well as a demonstration of a practical application on a commercial Swedish/English dictionary.

Chapter 10: A model for describing translation correspondences on structural and semantic levels is introduced.

Chapter 11: An application of the correspondence model from chapter 10. The model is applied to a sample of the Linköping Translation Corpus (LTC) and data on structural and semantic change are provided as well as more detailed information on some specific translation shifts.

Chapter 12: Conclusions, discussion and future work.

1.3 Relation to previously published work by the author

The contents of this thesis relates to previous work by the author and colleagues at the Department of Computer and Information Science as follows: The empirical study presented in chapter 3 is an extension of the work presented in Merkel (1998). Parts of chapter 4 and chapter 5 are extensions of the papers presented in Merkel, Nilsson and Ahrenberg (1994). The sections in chapter 6 relating to discrepancy analysis of LTC are an extension of Merkel (1996). Chapters 7, 8 and 9 are partly built on Ahrenberg, Andersson and Merkel (1998a, 1999), Merkel, Andersson and Ahrenberg (1999), Merkel and Ahrenberg (1999), and Merkel (1999). Finally, chapters 10 and 11 are a revised and extended version of Ahrenberg and Merkel (1997).

2 Background

In this chapter I will present some relevant background to the study, from the fields of corpus linguistics, translation studies, and computational linguistics.

2.1 Corpus linguistics and text corpora

The advent of "computer corpus linguistics", today known as corpus linguistics, dates back to 1961 when work on the text corpus later known as the Brown Corpus (Francis and Kucera 1964) was first being started. Linguists had traditionally been using the term "corpus" earlier than that to "designate a body of naturally-occurring (authentic) language data" as a basis for linguistic investigation (Leech 1997, p.1). Gradually the use of "corpus linguistics" has however become associated with language material which exists in electronic formats and the various methods and software tools that are used to analyse and access such data. The increasing processing power and storage capacity of computers have not only meant that the number of available text corpora has increased, but also that text corpora are larger in size, more varied and easier to access.

2.1.1 Application areas

Multi-lingual text corpora can be used as a source of information for several application areas, such as:

- Translation studies
- Contrastive linguistics
- Lexicography
- Computational linguistics and automatic translation.

Translation studies and text corpora will be described later in this chapter.

Recently a great deal of research within contrastive linguistics based on parallel and comparable corpora have been published, for example in Hasselgård and Oksefjell (1999) and Johansson and Oksefjell (1998). The availability of texts in electronic formats as well as designated tools for contrastive analysis have made it possible to conduct both more large-scale investigations as well as more focused studies on specific problem areas. The contrastive linguistics approach could both be applied to translation corpora (source and target texts) and to comparable corpora (which contain original texts in different languages).

Lexicography is also an area where empirical text material such as text corpora has been investigated. Langlois (1996) describes how a parallel corpus is used to develop a Canadian French-English and English-French dictionary with the aid of the TransSearch tool (Macklovitch 1992). The dictionary is due for publication in 2003. The COBUILD English Learner's and English Language dictionaries from Collins are monolingual and have been based on text corpora, which contain approximately 20 million words of running text (Sinclair 1987, 1995).

Within machine translation the cost of developing and maintaining translation lexicons is considerable. Therefore researchers have tried to use parallel corpora to extract lexicons with automatic methods in order to find translation candidates. In most approaches, the extracted translation candidates have to be validated by humans (Dagan and Church 1994) but recently new methods have been suggested where the accuracy of the automatic extraction methods could be increased to minimise the time the lexicographer has to spend on revising the translation candidates (Melamed 1998b).

Another branch of machine translation deals with the grammar and transfer rules in automatic translation systems. Parallel corpora prove to be a source of interest not only for lexical work, but also for finding patterns and constructions that constitute the basis for building grammatical and transfer components in MT applications (Grishman 1994, Kaji et al. 1992, Matsumoto et. al 1993, Meyers et al. 1998).

The above list is, of course, not complete; the use of text corpora has also great potential in for example foreign language teaching, literary studies and historical linguistics.

2.1.2 Types of text corpora

Before discussing different types of text corpora it is necessary to make clear what a text corpus is. In Mona Baker's words the original meaning of corpus is "any collection of writings, in a processed or unprocessed form" (Baker 1995 p.224). However, with the growth of corpus linguistics, this definition has to be modified: "(i) *corpus* now means a collection of texts held in machine-readable format and capable of being analysed automatically or semi-automatically in a variety of ways; (ii) a corpus is no longer restricted to 'writings' but includes spoken as well as written text, and (iii) a corpus may include a large number of texts from a variety of sources" (Baker ibid.).

Corpora can be divided into different types depending on what type of texts they contain. For example, corpora can contain general language texts or texts from restricted domains, written or spoken language, contemporary or historic texts, writer specific texts (a certain novelist) or genre specific texts (newspaper articles, fiction, court hearings, etc.), geographical variants (British vs. American English) and monolingual, bilingual or multilingual texts.

One distinction for corpora types concerns whether the corpus contains texts in one, two or several languages, that is, the distinction between *monolingual* and

multilingual corpora. Both monolingual and multilingual corpora can furthermore be divided into *parallel* corpora and *non-parallel* corpora (see Figure 2).

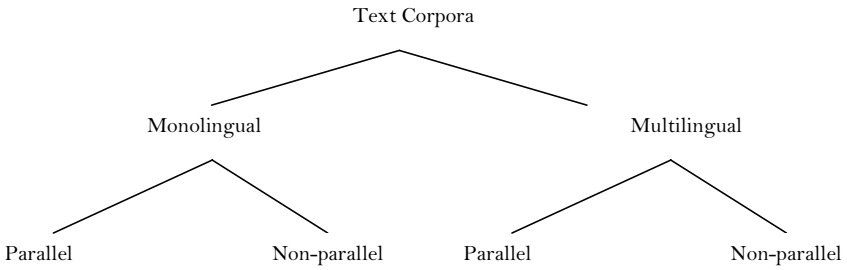


Figure 2 Types of Corpora

A monolingual corpus contains text(s) from one language and is probably the largest and most widespread type of corpus. Usually a monolingual corpus consists of running text, but not necessarily the complete text. The SUC corpus (Stockholm-Umeå Corpus) contains approximately 1,000,000 words and is comprised of 500 subtexts of 2,000 words each (SUC 1997). The Brown and LOB corpora also contain fragments of text of around 2,000 words each (Hofland and Johansson 1982). The British National Corpus is a considerably larger corpus, containing about 100 million words with 4,124 text fragments of around 45,000 words each (British National Corpus 1999). Parallel corpora within the same language (i.e., monolingual parallel texts) may seem slightly paradoxical, but are interesting and feasible objects of study, for example when comparing texts in different dialects, diachronical studies and comparisons of spoken vs. written language.

A multilingual corpus contains texts from more than one language and could be parallel or non-parallel. A *parallel corpus* consists of a source text and its corresponding target text(s), which are segmented on subtext levels, for example, on the paragraph or sentence level. For translation studies, it is vital that the texts are not only seen as parallel but also that the texts are viewed as either a source text or a target text. To date the majority of work on parallel corpora has been done on bilingual parallel texts, although one could foresee a future where the interest in more than two languages increases. But if translations are the object to study, multilingual parallel texts must always be viewed as sets of bilingual parallel corpora that share the same source text. Of course, the different target texts could be compared, but then the focus would not be on translation, which must take into account the original text and its different realisations of texts in different languages.

It is clear that the majority of corpora today are monolingual, but more and more bilingual and multilingual corpora are being developed around the world.

The most well-known and used bilingual corpus is probably the Canadian Hansard Corpus (used by for example Klavans and Tzoukermann (1990), Gale and Church (1991), Macklovitch (1992) and Simard et al. (1993). Other bilingual corpora developed during the nineties are for example the PEDANT corpus in Gothenburg (Ridings 1998), the English-Swedish parallel corpus (ESPC) in Lund (Aijmer et al. 1996), the PLUG corpus in Linköping, Uppsala and Gothenburg (Ahrenberg et al. 1998b), the English-Norwegian parallel corpus (ENPC) in Oslo (Johansson et al. 1996) and the HKUST English Chinese Bilingual Corpora (Wu 1994).

Non-parallel multilingual corpora can be of interest for comparative studies, even though the different texts are not originals and corresponding translations. Instead certain text types can be studied cross-linguistically by comparing original texts in both languages, or, if translation phenomena are the object of study, translations can be studied. Often such corpora are regarded as *comparable corpora* (cf. Laviosa 1998, Altenberg 1998, Aijmer 1999). The English-Norwegian and the English-Swedish parallel corpora contain translations in both directions (from English and to English) which makes it possible to use the corpus both as a parallel (translation) corpus and as a comparable corpus.

In Table 1 the different types of text corpora are illustrated from the perspective of multilinguality and parallelism.

Table 1. Types of corpora (multilinguality vs. parallelism)

	Parallel		Non-parallel	
Mono-lingual corpus	Diachronic corpus	For example Chaucer's Canterbury Tales in Medieval English vs. modern English versions.	Original-translation corpus	Original text vs. translated text in the same language, e.g. Swedish novels and English novels translated into Swedish (Gellerstam 1996) or newspaper articles (Laviosa 1998).
	Transcription corpus	Transcriptions of dialect versions of a standard language text or phonetic transcriptions of spoken language.	Text type corpus (one genre)	Monolingual corpus containing texts from the same text genre.
	Target variant corpus	Different translations into the same target language of the same original text (cf. Platzack 1983).	Mixed text type corpus	Different text types in the same language, usually balanced.
Multi-lingual corpus	Translation corpus	Source text and target text.	Text type corpus (one genre)	Multilingual corpus containing texts from the same text genre.
	Multi-target corpus	Several target texts in different languages originating from one source text.	Mixed text type corpus	Different text types in several languages, usually balanced.
	Mixed source corpus	Several parallel texts where the original is unknown (cf. EU texts and certain Bible translations, e.g. Melamed (1998c)).		

Usually the texts in parallel corpora are linked (or aligned) on the sentence level. An example of a sentence-aligned text from a translation of a computer program manual into Swedish is shown in Table 2.

Table 2. Example of sentence aligned text

Source Text	Target Text	Mapping
Microsoft Access 2.0 has new built-in toolbars that you can modify to fit your requirements, move around in the Microsoft Access window, and hide or show individually.	Microsoft Access version 2.0 har nya, inbyggda verktygsfält som du kan ändra så att de passar dina speciella krav. Du kan flytta runt dem i fönstret för Microsoft Access och gömma eller visa dem individuellt.	1-2
You can also create your own custom toolbars.	Du kan också skapa egna verktygsfält.	1-1
To hide or show an individual toolbar, including custom toolbars you create yourself, choose Toolbars from the View menu.	När du vill visa eller gömma ett visst verktygsfält, även sådana som du själv skapat, väljer du Verktygsfält på Visa-menyn.	1-1
You can also hide or show a toolbar in a macro or module by using the ShowToolbar action.	Du kan också visa eller gömma verktygsfält från ett makro eller en modul med instruktionen Visa verktygsfält.	1-1
To hide or show all built-in toolbars, choose Options from the View menu and in the General category, set the Built-In Toolbars Available option (similar to the Show Tool Bar option in version 1.x).	När du vill visa eller gömma alla inbyggda verktygsfält väljer du Alternativ på Visa-menyn. I kategorin Allmänt anger du inställningen Ja eller Nej för Visa inbyggda verktygsfält. Detta motsvarar alternativet Visa verktygsfält i version 1.x.	1-3

The column "Mapping" shows the sentence relationship between the source and target segments. The first English sentence has been translated with two Swedish sentences and the last English sentence with three Swedish sentences.

The concept of "bilingual parallel text" is derived from Harris' (1988) introduction of the *bitext* concept where the two texts are connected on different segment levels into a whole bitext, "stored in such a way that each retrievable segment consists of a segment in one language linked to a segment in the other language which has the same meaning" (Harris, p. 8-9).

As will be shown later (see Figure 4), the fact that translations can be made with different methods, with different objectives concerning the relationships between the source and target text, means that parallel translation corpora will contain different types of parallel texts. Hartmann distinguishes between three types of parallel texts (Hartmann 1997):

1. Texts that are the result of a full-scale translation act (e.g. novels)

2. Texts that are the result of interlingual adaptation (e.g. advertisements and multilingual formulations of documents published by international organisations)

3. Texts that are not translationally equivalent, but functionally similar in situationally motivation and rhetorical structure (e.g. cooking recipes, wedding announcements).

Type 2 might actually be considered as a subtype of 1 and this merged category of parallel texts has been termed "bitext" by Harris (1988), meaning "a source text and target text as they co-exist in the translator's mind at the moment of translating". Hence, a bitext or a parallel text does not distinguish between a pure translation (where each unit is translated into a more or less literal target) and an adapted text (where the target text cannot be literally translated). Consider the following example from a software manual translated from English into Swedish:

> ***Microsoft Excel Visual Basic Reference*** A complete reference to the Visual Basic language is available in online Help. This reference is also available in book form and can be purchased in book and software stores or ordered directly from Microsoft Press. To place a credit card order, call 615-793-5090, or toll-free 800-MS-PRESS. Be sure to have your reference code FXL ready for faster order processing.
>
> ***Microsoft Excel Visual Basic Referens*** En fullständig beskrivning av Visual Basic-språket finns i direkthjälpen. Den här referenshandboken finns även i bokform och kan köpas i bokhandeln eller i programvaruaffärer.

The more elaborate description of how to order the book in question directly from the publisher has been omitted in the Swedish translation of this paragraph, because it has been *adapted* to a typical Swedish user who will not or cannot call toll-free numbers to the United States.

In most industrialised types of translations it would be very difficult to find a completely "purified" kind of translation (read literal). Some degree of adaptation is needed in this kind of translation, and although stretches of text may be considered as belonging to Hartmann's type 1 above, there will certainly be passages where adaptation is preferred and indeed absolutely necessary.

2.2 Translation studies

The field of translation studies has not yet reached the stage where researchers are willing to label it as translation *science*. It is notable that two of the most influential works within the field express this in the book titles: Eugene Nida's "Toward a Science of Translation" (Nida 1964) and Gideon Toury's "In Search of a Theory of Translation" (Toury 1980).

Translation studies, as the field is most commonly referred to, is an empirical discipline and should according to Toury (1995) be devised to describe and explain certain segments of phenomena of language and language use in the "real world". Like all branches of science, translation studies should contain a descriptive branch where phenomena on the object level can be described, explained and predicted.

James S. Holmes' (1972) description of the translation studies field is depicted in Figure 1:

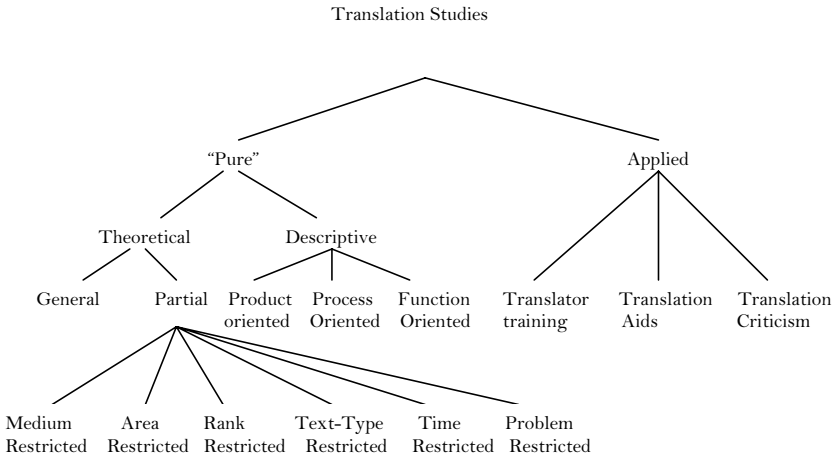


Figure 3. Holmes outline of Translation studies (in Toury 1995, p. 10)

The major division lies in Holmes view between the "Pure" and Applied branches. Pure translation studies are broken down into a theoretical and a descriptive branch which in turn are divided into the general and partial sub-branches for the theoretical side, and product, process and function-oriented for the descriptive branch. On the applied side of translation studies we find the more concrete manifestations of translation work, such as training of translators, translation aids and translation criticism.

2.2.1 Product-, process- and function-oriented approaches

The division of Descriptive Translation Studies into product-, process- and function-oriented approaches is of major importance. In a *product-oriented* approach, it is the concrete products and results of translation work that are the object of study, namely the source and target texts. These texts are the empirical basis for any product-oriented investigation. A *process-oriented* approach has its focus on *how* a translation is created, what decisions a translator makes, what strategies she uses, etc. In a *function-oriented* approach to descriptive translation studies the research is aimed at the recovering the function of a translation, the role and position of the translation in the target culture and target context.

The most commonly used method of studying the process of translation is by using Think-Aloud Protocols (TAPs), for example, TAPs have been used by Dechert and Sandrock (1986) and Lörcher (1992). The think-aloud protocols are used to collect introspective data; basically the subjects are asked to state aloud what comes to mind while translating. For example, TAPs have made it possible to test whether translators find it more problematic to translate *into* their mother tongue or *from* their mother tongue (Lörcher 1992).

The role and function of the translation will certainly determine which strategies a translator uses during translation and therefore also what characterises the translation process. A product-oriented approach can therefore not be taken in isolation from "questions pertaining to the determining force of its intended function and to the strategies governed by the norms of establishing a 'proper' product" (Toury 1995, p. 13). In other words, even if the main objective is to study the object relationships between an original and its translation, one must also take into account in what context the translation is going to be used (for example, as a computer manual for novice users) or what possible constraints there could be for transferring relevant information (or attitudes) into the target setting.

The approach taken in this thesis is dominated by a product-oriented viewpoint; it is the source texts and their translations that are the principal objects of study. Moreover, the study has a specific application in mind, namely translation aids, so the findings are primarily judged for their relevance to this application. However, linguistic properties and constraints of the source and target language must also be considered as well as text type, availability of resources, number of translators involved, etc.

In the ideal situation, the field of translation science should be able to let theory and empirical results form the basis for the applied translation fields taught at universities and, in particular, the different tools and aids that are produced to improve translation work. The research philosophy initiated by Pierre Isabelle and the group at the University of Montreal adhere to this basic idea too, as evidenced by the quote in the Introduction, page 1. The assumption is of vital importance in this work; practical translation tools and translation methodology should be firmly based on empirical data present in existing translations.

2.2.2 Types of translation

As has been pointed out by Baker (1993), translated texts play a very important role in forming people's experience and knowledge of other cultures. Most of the famous works in the world literature come to us only in translated versions. Furthermore, we are dependent on translations of news bulletins, legal documents, operating manuals for technical appliances, and various other central areas in our lives.

> Given that translated texts play such an important role in shaping our experience of life and our view of the world, it is difficult to understand why translation has traditionally been viewed as a second-rate activity, not worthy of serious academic enquiry, and why translated texts have been regarded as no more than second-hand and distorted versions of 'real' texts. (Baker, ibid. p. 233)

For small languages like Swedish the influence of translation on written text in general is more significant than for world languages like English, Spanish and French. Before the middle of the 18th century, translation dominated the production of text in Swedish. Furthermore, the majority of literary text published in Swedish in the last two centuries consists of translations (Wollin 1993).

During the sixties and seventies translation studies were heavily source-oriented, that is the main focus was on the source text and its features. The main objective was to preserve and protect the "legitimate rights" of the source text in translation (Toury 1995). Features and constraints present in the target language were considered of secondary importance. However, with the publications by Vermeer (1978), and work by Toury (1977, 1980) a gradual shift towards target-oriented translation studies started. Today, especially in applied translation studies, such as Newmark (1988) and Ingo (1991) the target-oriented approach is the prevailing way to approach practical translation work.

The term *translationese* has been used to describe the "systematic influence on target language (TL) from source language (SL)" (Gellerstam 1985). Gellerstam reported that English source language text has had a considerable effect on the Swedish translations. Although the translations are grammatically correct, the traces of the source text are visible in the target text, for example, when you compare translated novels with novels written in the original language. Certain constructions and lexical items will have a much higher frequency in the translated novels than in novels written originally in the same language. An illustrative example from Gellerstam's study of Swedish translations of English novels is the verb *anlända* which is a standard translation of the English *arrive*, but in original Swedish fiction this verb is hardly used at all (instead the verb *komma (till)* is the preferred choice). Furthermore, tag questions like "eller hur" have a much higher relative frequency in the translations compared to original Swedish fiction. The existence of translationese in target texts may merely be a result of unprofessional translators[1], the mechanical nature of second language training in schools or a gradual, almost imperceptible, influence from a dominating world language. The use of translationese may on the other hand be a conscious choice by the translator in order to achieve a certain effect.

The way that a translation is related to its source and target text has been described by for example Newmark (1988). A translation can be placed, in a general sense, on a continuum from being source language-oriented or target

[1] In the Swedish newspaper Dagens Nyheter a literary critic draws attention to a piece of poor translation work of a recently published book on March 15, 1999. The translator had translated the English "Sleeping Beauty" with the translationese "sovande skönhet" instead of the standard proper name from the fairy-tale:"Törnrosa".

language-oriented. This continuum of translation orientation is illustrated in Figure 4 below.

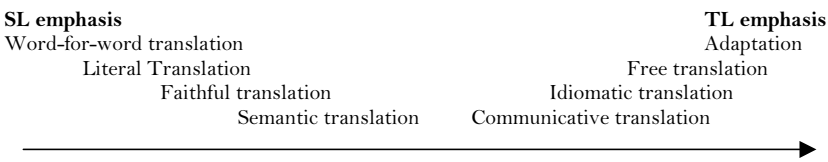


Figure 4. Different translation methods (Newmark 1998, p. 45)

The *word-for-word translation* method is usually used for illustration purposes or as a stage in the translation process where each source language segment is looked up word by word. Here grammatical issues in the target language are ignored and the word order in the target language mimics the order in the source language.

In *literal translation* the grammatical structure are preserved as closely as possible to their nearest grammatical TL equivalents. The lexical items are however translated out of context and one by one.

Faithful translation means that the translator tries to reproduce the exact contextual meanings of the SL text within the constraints of the TL grammar. Here the grammar and lexical choice are preserved in the target in a way that deviations from the SL norms present in the source text are transferred to the target text.

Semantic translation is similar to faithful translation, but with the difference that the aesthetic value of the SL text must be accounted for (e.g. "beautiful" sounds) and thereby compromising the content to a certain degree. Newmark writes that the difference between "faithful" and "semantic" translation "is that the first is uncompromising and dogmatic, while the second is more flexible".

Adaptation is the freest type of translation, usually found in plays and poetry where themes, character and plots are preserved, but little else.

In *free translation* the translator reproduces the content without adhering to the form of the original. Often it is a paraphrase that is much longer than the original.

Idiomatic translation is similar to free translation, but with the difference that there is a closer mapping between the form of the source and target texts. The nuances of meaning are however often distorted by colloquialisms and idioms in the translation, which are not present in the original.

Finally, by *communicative translation* the translator tries to find the exact contextual meaning of the original in a way that both form and content are accepted and understood by the intended readers. In short, this is the golden compromise between content and form.

Semantic translation is in Newmark's opinion personal and individual and follows the thought processes of the SL author. Communicative translation has a stronger focus on the target language. Or to put it in Newmark's own words "semantic translation has to interpret, a communicative translation to explain"(ibid. p.48).

2.2.3 What is "translation"?

It is difficult to find a definition of *translation* that most scholars agree on. Catford defines translation as "the replacement of textual material in one language (SL) by equivalent textual material in another language TL" (Catford 1965, p.20). Newmark's definition of translation is "rendering the meaning of a text into another language in the way the author intended the text" (Newmark 1988, p. 5). Lindquist (1989) offers a slightly more explicit definition of translation as a product when he states that a "translation is a text which is

(a) based on another text (the SL text) which is written in another language /.../

(b) written with the purpose of making some aspect(s) of the source text available to a new readership

(c) made in such a way and stands in such a relation to the source text that members of the target culture accept it as a translation". (Lindquist 1989, p. 18)

Lindquist's definition stresses the fact that a translation always is secondary in relation to the source text (point a). Point (b) is in a way contradicting both Catford's and Newmark's definitions because of Lindquist's claim that it is not necessary to convey *all* aspects of the source text in the target text. The reason for the weakening point b is that it may not be possible to include certain aspects of a source text in a translation. There are many translations, often of literary works, that do not contain all the features and aspects present in the source text, simply because of the fact that some of them cannot be translated without sacrificing style and readability. The last point stresses the fact that translations are artefacts of the target culture. Over time the demands from the target culture will change, which means that a piece of text that was once a perfectly acceptable translation, for example, in the 18th century, will not be accepted by the present-day target culture.

One particular interest for this thesis lies in the second point that Lindquist makes, namely, that not all aspects of a source text are always translated. Is it possible to explain and predict when something is left out? And furthermore, sometimes a translator *adds* items in the target text that may have been implicit in the original, but not stated explicitly.

2.2.4 Correspondence

A central issue in translation studies has been dominant until recently: the notion of *equivalence*. Translations should be as equivalent as possible in relation to the original, (equivalence should here be understood as a semantic or formal

category). The focus has therefore been "to determine what an ideal translation, as an instance, should strive to be in order to minimise its inevitable distortion of the message" (Baker 1993, p. 236). The term "equivalence" also has the connotations of taking a qualitative stand on a certain translation as it is also regarded as focusing on the *relationship* of the corresponding source and target constructions (Ingo 1991). The notion of *translation correspondence* is more neutral in this respect since it can be used to describe translations that are mediocre, bad or even erroneous, as well as good or acceptable translations.

Haas (1968) was a pioneer in his opposition to the traditional equivalence view by taking a more situational stand. Two sentences are equivalent (in meaning) if "there is a correspondence between their uses" (ibid. p. 104). Haas advocated a shift from equivalence of meaning to equivalence of usage. This also implies that it is possible to take a wider look at the context in which the segment being translated occurs and view the context as the situation which will play an important role in the way the segment in question is being translated.

Translation involves the application of *translation operations* to achieve the intended correspondence between the source and target text. A number of translation operations are described in Newmark (1988, pp. 81-88). The operations discussed here are *transference*, *naturalisation*, *cultural* and *functional equivalence*, *through translations* and *shifts*.

(a) *Transference* indicates the process of transferring a SL word to a TL text. The TL word becomes a loan word, with the SL spelling and form. For example, the English use of the phrase "The Red Lion" when used about a pub can be transferred into Swedish unchanged as "The Red Lion".

(b) *Naturalisation* is transference plus adaptation of the SL word to the normal pronunciation and morphology of the target language. For example, the English word "thriller" has been naturalised into Swedish (with Swedish pronunciation and Swedish inflectional patterns).

(c) *Cultural* and *functional equivalence* (Nida 1964) entails finding approximate equivalents in different cultures. The English "A-levels" may correspond to the Swedish "gymnasieexamen" in certain contexts. Sometimes the use of this procedure neutralises or generalises the SL and sometimes it adds something. For example the English "Roget's" may be translated as "engelsk synonymordbok".

(d) *Through translation* (also referred to as *calquing*) is the process that produces translationese (see the discussion earlier in this chapter). It involves the literal translation of common collocations, names of organisations, (Fr. *compliments de la saison* - Eng. *compliments of the season*).

(e) *Shifts* (or *transpositions*) involve a change in the grammar from SL to TL. Sometimes shifts are obligatory since the SL structure may not exist in the TL. Sometimes the literal translation is possible, but it does not create a natural TL passage so the shift is necessary to arrive at a satisfactory translation. In other cases, lexical gaps are filled by a grammatical structure in the TL.

All of the above operations can be found when translations are studied in detail. According to Toury (1995), it is by comparing the source texts to the target texts that the translation relationships can be uncovered and generalisations about the underlying concept or norm of translation can be made. The implications are that a deeper understanding of the concepts of translation would help to influence future translations and future studies of translation. One of the questions in focus for this work is to find out more about the relative presence of different translation shifts from the original to the translation (see chapters 10 and 11).

2.2.5 Universal features of translation

An important area for corpus linguistics and translation studies could be to identify universal features of translation, features which typically occur in translated texts as opposed to original text and where these features "are not the result of interference from specific linguistic systems" (Baker 1993). Examples of such universal features according to Baker are:

1. Higher level of explicitness in the target text

2. A tendency towards disambiguation and simplification.

3. A strong preference for conventional 'grammaticality'

4. A tendency to avoid repetitions which occur in source texts

5. A general tendency to exaggerate features of the target language, certain constructions common in the target language tend to occur more frequently in translated material rather than in original material in the target language.

6. A "translationese" tendency, certain lexical items and syntactic constructions have a higher frequency in translated texts than in texts of the same genre in the same language because of source language interference.

Features 5 and 6 seem to be contradictory, but in Baker's argumentation these features could coexist. From the universals listed above, we should expect that translations are more specific, clearer, more grammatical, less repetitious and involve a degree of translationese. To be able to verify the existence of such universals, large translation corpora and appropriate analysis tools must be available.

2.2.6 Technical translation and the technical translator

Good technical language is usually free from emotive language connotations and original metaphors. However, the most crucial part of technical translations is the use and handling of terminology, especially new terminology (Newmark 1988).

The translation style may vary according to the requirements of the customer. Some customers do not issue concrete stylistic guidelines to the translators; instead the translators have to use their acquired experience and expertise to

make their own judgements. On the other hand, some customers have detailed translation guidelines for each language, both for terminological and stylistic aspects of the text. For Swedish, Microsoft, IBM and Apple have for example produced their own translation styleguides (Microsoft 1993, Ström and Windfeldt 1991, Apple 1994).

Hann (1992) describes the technical translator as falling into one of four different types. The first is the "dictionary enthusiast" who invests in dictionaries and looks up the same term in different dictionaries to single out the most frequently occurring translation as the correct one. The second type uses dictionaries sparingly and prefers to get first-hand information from encyclopaedias and thereby identify concepts in the two languages directly. The third type of translator consults standard textbooks on different subject fields in both languages and tries to identify equivalent collocations by making use of the respective indexes. Finally, there is the fourth type of technical translator who relies on computers, often more than one. He prefers electronic term banks and access tools rather than printed dictionaries and may compensate a possible lack of expertise in engineering by skill in manipulating information systems.

With the evolving availability of computerised translation aids, we will see the fourth type of technical translator emerging as the dominant one. There are already computerised translation aids, such as multi-lingual term-banks on specific technical domains and encyclopaedias on CD-ROM discs that can replace the printed versions of these information sources for types 1 and 2. The third type of activity, to consult textbooks in the respective languages, may not develop as fast as the first two in the direction towards electronic formats, but many technical magazines and publishing houses do make their texts available on CD-ROM discs as well.

2.2.7 Translation of fiction

In the translation of serious literature the emphasis is both on the form and the content of the message. The artistic qualities present in the SL text must be preserved or recreated in the TL with minimal distortion of content. It is notoriously difficult to give general characteristics for literary translation, but the relative impact of the SL culture and the moral purpose of the author to the reader may surface in how proper names, dialogue, dialect, non-standard language, etc., are translated.

2.3 Text corpora in translation studies

With the emergence of more powerful computers, more readily available text on electronic formats and new software aimed at analysing text, the field of translation studies has entered a new stage. Many researchers have advocated the use of corpora for various purposes: for training translators (Lindquist 1984), for studying "translationese" (Gellerstam 1985, 1993, 1996) and for developing translation theory (Baker 1995) and Toury (1995).

To be able to examine and analyse a text corpus the researcher need a set of corpus tools (cf. Atkins and Clear, 1992). This set of basic tools for a *monolingual corpus* can include:

Frequency tools: Tools that produce lists of word types with their frequency in the corpus.

Concordancing tools: Text retrieval and indexing software, which provides the possibility to search and examine linguistic contexts.

Lemmatization tools: Software that relates an inflected word form to its base form or lemma.

Parts-of-speech tagging: Software that assigns a word class label to every word in the text.

Syntactic parsing: Software that assigns a full or partial syntactic constituent tree to sentences in the text.

Search tools: Software that can search for various patterns present in the corpus, depending on the format of the corpus (base forms, parts-of-speech patterns, surface forms, etc.)

Collocation retrieval: Software that extracts collocations and multi-word units from the text.

For work on a *parallel corpus* the above mentioned tools are naturally of use for analysing either the source text or target text separately. However, to capture the relationships *between* the original and the translation, a translation scholar needs to extend the tool kit. The following set of tools is needed to set up and process parallel corpora:

Paragraph and sentence alignment tools: Software that links paragraphs and sentences into corresponding units. Sentence alignment is the usual prerequisite for any work on parallel corpora as it is with these alignment tools that the parallel texts are created.

Bilingual concordancing tools: Concordancing tools that process a parallel corpus, which means that when a source word is looked up in the source text, all contexts where the given word occurs, including the corresponding target sentences, are presented to the user.

Word alignment tools: Software that creates correspondences between units below the sentence level, usually one-to-one relationships between words but also between multi-word units.

Parallel search tools: Software that makes searching in both the source and target text possible, for example the user could search for all source-target sentence pairs where a certain source word co-occurs with a certain target word.

In addition to the above mentioned tools, there are various other functions that the translation scholar could have use of, for example tools that compute various statistics measures from the parallel corpus, such as type-token ratios, length relationships between source and target texts, translation consistency, translation discrepancies, etc.

2.4 Translation and computational linguistics

Scientists have tried to apply computers to translation ever since the Second World War, with varying degrees of success. The main reason has been a practical one, "in the era of information explosion translation becomes a very critical business" (Nirenburg 1987, p. 1).

2.4.1 Machine Translation

Back in the fifties fully automatic translation was a desirable goal in order to handle the information flow and to enhance efficiency. It was also believed in the early days that it would be relatively straightforward to specify and formalise the tasks involved in translation and thereby be able to construct translation systems. Nirenburg writes that the belief was that "since translation is such a common everyday task, performed with relative ease by humans, it must be easy to automate" (ibid. p 1). This belief was predominant in machine translation research up till the mid sixties. Hutchins (1986) describes MT under this period as "an area of intensive research activity and the focus of much public attention". This intensive period came to an end after the publication of the ALPAC report in 1966 where it was stated that machine translation was definitely not a solution to the above mentioned problems. It was reported that "there has been no machine translation of general scientific text, and none is in immediate prospect".[2] In the ALPAC report it was also stated that all MT output had been post-edited and this was seen as a failure. As Hutchins points out, they failed to recognise that human translators also revise their translations before they submit them to their clients. From the setback in the sixties efforts within MT were relatively small until the revival in the late seventies. In Europe the MT renaissance was shown by the fact that the Commission of the European Communities purchased a version of the English-French SYSTRAN system and they also commissioned the development of other language versions. The EUROTRA project was founded where the aim was to create a "pre-industrial" MT-system for the EC languages (Arnold et al. 1994). In Japan several MT projects were initiated with novel approaches to MT. In the U.S. and Canada there were also examples of the budding MT interest: the TAUM group at the University of Montreal started their work which led to the METEO system (Chandioux and Grimaila 1996) and the U.S. Air Force funded work on the METAL system at the Linguistics Research center, University of Texas, for instance.

On the Swedish arena, research groups in Uppsala and Lund have been active within MT. In Uppsala an MT prototype system called Multra (multi-lingual

[2] Quoted from Hutchins (1986, p. 165)

support for translation and writing) has been developed. Multra is a transfer-based system and has transfer components for Swedish-German and Swedish-English (Sågvall Hein 1994). In Lund different versions of the SWETRA system have been adapted to domains such as weather forecasts and stock market reports (Sigurd et al. 1992). The SLT system (Spoken Language Translation) originally developed at Stanford Research Institute has been modified to handle Swedish/English at the Swedish Instititute of Computer Science (SICS) and Telia Research. SLT is also a transfer-based system that can translate queries about air travel (Agnäs et al. 1994).

On the commercial side there are no advanced MT systems available for the Swedish language. As far as I can tell in the autum of 1999, there are two word-to-word translation programs around: *Tolken* and *Engelska hjälpredan*.[3] Even though the systems contain rather large lexicons, the techniques used are primitive and without grammatical analyses. The translation quality from Engelska hjälpredan has been evaluated by Granlund (1999) who experienced problems with word order, lexical selection, agreement and tense. Although these two systems are not meant for serious commercial translations, they may help users to create rough translations.

Systran has developed a prototype version for English/Swedish, but so far it has not reached the market. The Swedish prototype is not as advanced as the commercial versions for other languages. The shortcomings are mostly connected to the Swedish lexicon and the possibilities for updating.

In the 1990s the area of computers and translation has been firmly established, both as commercially available systems and as an accepted area of research. The problems of successful MT are by no means solved, but the ground has been set with a more realistic outlook on what computers can do for translation, compared to the expectations that prevailed in the fifties.

There are attempts at more empirical methodologies for MT, namely example-based and statistical MT. Most of the commercial and research MT systems are *rule-based* systems, i.e. they rely on on syntactic and semantic representations, which have to be specified for the systems. In *example-based MT* systems *analogy* is taken as the guiding light. By using parallel corpora, patterns and templates are derived from the source and target texts, and used by the translation systems. The first suggestion to this approach was made by Nagao (1984). Other example-based approaches can be found in for example, Sato and Nago (1990), Jones (1992), Kaji et al. (1992), Sumita et al (1993) and Foster et al. (1997). In statistical MT approaches, explicit rules are replaced by statistical methods to process bitexts (e.g. Brown et al. 1988). By using the bitext and a statistical translation model, the possible translations for the words of the source language can be identified. The target words that are suggested for a given sentence are then ordered on the basis of a target language model that generates the statically most probable ordering between the proposed words. Hybrid approaches where rule-based, example-based and statistics-based MT are combined have also been proposed, for example, in the Pangloss III system (Brown 1996).

[3] Tolken can be downloaded at http://www.algonet.se/~hagsten/ and Engelska hjälpredan at http://www.internetami.nu/ (September 1999).

The automatic translation systems that produce text without intervention by humans are being accompanied by systems where the translator is in focus or at least has the possibility to influence or guide the computer in its translation process.

The approach to keep the translator at the heart of the translation process and to let her control various sources of information and translation tools stands in sharp contrast to the machine-oriented translation view where a computer system is the primary agent and the human translator/editor is only active in preparing the system and revising the output of the system. Fully Automated High Quality Machine Translation (FAHQMT) is more of a dream than a realistic option for most types of texts and language pairs. Only in very limited domains with a high degree of repetitive and standardised language can FAHQMT become a path to tread. But if the requirements on quality can be lowered, then automatic translation may be very useful, for example, when you want to get rough translations which is only intended for users who want to get the gist of a foreign text or to decide if the text in question is interesting enough to have it translated with a quality that meets to publishing standards.

2.4.2 The Translator's Workbench

Almost twenty years have passed since Martin Kay wrote about the relationship between computers, translation and translators (Kay 1980). In his report, which was never published but nevertheless often quoted, Kay expressed his ideas on how computers could be used for translation work, given the contemporary state, and relative failure, for the application of machine translation in *real* translation activities. Kay's basic ideas were that it would be unrealistic to expect any major breakthrough for high quality machine translation within a foreseeable future and that one instead should adopt a more modest approach to the role of the computer in translation. The translations produced by automatic translation systems cause more problems than they help. Translators (or post-editors) have to spend more time repairing the damages done to the text by the MT system than it would take to translate it manually from scratch. Kay's proposal was instead to let the computers take care of things that are difficult or time-consuming for human beings (such as looking up words in a dictionary or searching through a bank of old translations) and let humans do what computers have difficulties in (deciding what the appropriate way of phrasing a certain passage in the target language is). The suggested machinery was in Kay's words called *The Translator's Amanuensis*. The idea was to see the problem of how computers should be used in translation as a gradual process where computers "almost imperceptibly" are "allowed to take over functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!" (ibid. p. 13).

Many people in the field seem to whole-heartedly agree with Kay's proposal, but there is still, almost 20 years later, no agreement on what these tools ("functions") are. Isabelle et al. (1993) discuss these issues and suggest that there should be a distinction between *office automation for translators* and *translation*

support tools. In the former category they mention functionalities such as split-screen word processing, spelling correction, terminology and dictionary lookup, file comparison, word counting and full-text retrieval. As a foundation for translation support tools they claim that *translation analysis* is a key notion, in which applications such as translation memory, translation checking and a translation dictation machine fall. Church and Hovy (1993) also try to define tools that would agree with Kay's proposal and they suggest *super-fast typewriting* (where words are completed automatically when enough of the word has been entered to make it unambiguous), *bilingual concordances*, and *raw translations* (in "cliff-note" mode, i.e., automatic dictionary lookup of words and phrases) for email messages, etc.

In 1997 a whole issue of the journal Machine Translation (vol. 12 Nos. 1-2) was devoted to new tools for human translators where the starting point was the reprint of Martin Kay's paper from 1980. The question raised in the preface of the issue was to look at what had happened in the MT field and to the proposed translator workstation approach.

Today Kay's criticism of MT can still be seen as valid. Although there are many more MT systems available today than twenty years ago, the basic technology behind the systems remains the same. Several of the companies that were around twenty years ago have downsized their mainframe systems (Systran, Metal and Logos) and retained most of the features of the old systems in versions for the personal computer (Hutchins 1996). They are cheaper, available to many more people and a commercial success for the companies that produce them (Flanagan 1996, 1997). For many users, it is not even necessary to have your own translation system today, the products can be used over the Internet for free, (for example Systran's translation system through the AltaVista search engine (Yang and Lange 1998).

Kay's vision of the Translator's Amanuensis has in part come true; the increased functionality in word processors, the development of on-line lexicons and, perhaps most notable, the dedicated pieces of software called Translator Workstations, which are based on the concept of Translation Memory, (for example, Trados' Translation Workbench, IBM's Translation Manager, GlobalWare's XL8 and STAR's TransIt). Translator Workstations can be seen as translation environments which are compatible with word processors and which contain features that aim at increasing the productivity of translators, such as translation memories, split-screen word processing (for the source and target text), terminology look-up and management, etc.

Translation memory (TM) systems are built on the principle of recycling. During the translation process a database of corresponding source and target sentences (translation memory) is created incrementally. Anytime the translator comes across a translated passage in the source text, the translation memory tool will present the previously made translation as a suggestion that the translator can accept as it is or revise it. The TM systems also include some kind of fuzzy matching, which means that source sentences that are almost identical will also be found and presented to the translator. The optimal use of translation memories comes when different versions of source text have to be translated or

when, for example, technical manuals are updated with only minor changes to the original text.

In the localisation industry, translation memory tools are nowadays an accepted part of the translation work (O'Brien 1999). Many customers require that translations are delivered from the translation agencies as translation memories. The TM- tools have therefore put new demands on professional translators, both freelance translators and translation companies. Not only do they have to invest in new software and training, there is also the additional need for maintaining and administering the growing archives of translation memories. For many the translator workbench idea as it was realised in TM packages has failed to live up to the initial optimistic expectations. O'Brien states that the "biggest failure of Translation Memory has been its inability to deliver on the expected cost and time reductions" (O'Brien 1999, p. 8). This hesitation towards the promises of productivity gains from the software developers was already expressed in 1994 by Schäler (1994). Some of the encountered problems have to do with the "sharing of translation memories". In many cases, especially when a large team of translators work with TM tools, it can be difficult to organise the translation memory in such a way that it can be efficiently shared over a network. This may have to do with the involvement of freelance translators who are not physically in the same building as the in-house translators and project management. It can also depend on technical issues, e.g. the capacity of the local network, etc. But maybe the most serious problem stems from the fact that different versions of one TM can be floating around a company and sometimes these TMs are merged into a "new" one without proper validation. This could result in excellent translations being overwritten by not so good translations.

Another issue that concerns quality and the effects of TM tools on the translations has been admitted by the TM software developers themselves (Heyn 1998). One of these issues involves the "peep-hole effect" that arises when TM tools force the translator to focus on one translation segment (usually a sentence) at a time. The active translation segment will occupy a large part of the screen and might cause the translator to spend less time connecting the translation to the surrounding context. Pronouns and coherence markers that bind the text together can also be omitted in order to make the translation as recyclable as possible in other surroundings. The effects of using translation memory tools are addressed in several places in this thesis, in chapters 3, 6, (section 6.2) and chapter 11.

The revived interest for corpus-based approaches in NLP research has produced improved methods and approaches for translation tools. The necessity to have access to (and to be able to create) parallel corpora (a source text and its translation) has caused many research groups to develop and improve techniques and methodology for corpus work (such as alignment programs and different ways of extracting linguistic data from electronic texts).

Within the research fields, focus has shifted from the pure MT approaches; i.e. exploring and developing new approaches to MT to activities which are in line with the Translator's Amanuensis approach. Within computational linguistics, techniques for building reusable resources (such as parallel texts, bilingual

dictionaries, and extractions of collocations and terms) have in recent years caused considerable more interest than exploring the possibilities of fully automatic MT.

There is also a dividing line between the MT systems camp and the Translator Workstation camp in geographical terms. In the United States and Japan, the commercial interest for MT packages has exploded from 1994, especially for personal MT packages whereas the MT interest in Europe has been limited to large translation service providers and multinational companies (Flanagan 1996, Hutchins 1996). The European focus has instead been on the Translator Workstation approach. It is worth pointing out that all leading producers of Translator Workstation software are European.

2.5 Building parallel corpora – sentence alignment

In order to reuse translated material, there must be tools and methods to build resources that can be utilised during translation. One such example is to construct bitexts from existing translations. This is called *aligning* the source and the target texts. The most investigated area of alignment is performed on sentence level, but it is also possible to align both larger segments (such as paragraphs) and smaller segments (such as phrases and words).

There are two main approaches to sentence alignment, namely length-based alignment and lexicon-based alignment. The length-based alignment approach can be based on either character length (Gale and Church 1991) or word length (Brown et al. 1991). The lexicon-based alignment is used by for example Mariani et al. (1991) to produce correspondences between words by means of a bilingual lexicon. Hybrid approaches to alignment have also been proposed by, for example, Johansson and Hofland (1994) and Wu (1994), where statistics and bilingual lexicons are used together. Here the idea is that the use of a list of pairs of source and target words with a high likelihood of being translations complements the statistical alignment algorithm. In Johansson and Hofland's terminology this list is composed of *anchor words* whereas Wu calls the same *lexical cues*, but the basic functionality is the same. Chen (1993) used a slightly different approach when he used lexical information for sentence alignment by adding a statistical translation model to the Brown approach (Brown et al. 1991).

The general formulation of the statistical approach can be described as to "choose the alignment that maximises the probability over all possible alignments" (Wu 1994, p. 81). But this is too general to be of any practical use and it is instead the different ways of implementing the approximations that have been controversial.

Brown et al. (1991) proposed a length-based alignment algorithm where the number of words in the sentences were taken as the primary criterion. They applied their algorithm on the English-French Hansard corpora and report a high success rate in aligning sentences. Gale and Church (1991) showed that it is better to measure length in terms of characters than words, especially when you compare languages that have different characteristics as regards compounding.

Corresponding sentences are less likely to vary in numbers of characters compared to words. Gale's and Church's algorithm has been very influential and forms the basis for several other proposals.

Accurate sentence alignment must be able to account for different mapping relations between sentences in the source and target text, such as 1-0 (sentence deletion), 0-1 (sentence insertion) 2-1 (sentence combination of two source sentences into one target sentence), 1-2 (sentence splitting where 1 source sentence corresponds to 2 target sentences), 2-2 (two source language sentences correspond to two target language sentences). There are other possibilities of combination and splitting relations, such 1-3, 3-1, 3-3, etc. However, the most commonly observed mapping relations are 1-1, 2-1, 1-2, 1-0 and 0-1. An example of combination and splitting mappings is shown in Table 3.

Table 3. Complex sentence mappings

Source	Target	Map	
"Yes, with one hand while you were busy stirring a pot with the other.	"Ja, med ena handen. Medan du rörde i grytorna för fullt med den	MAP 1-2	(From: Gordimer, A Guest of Honor)
"Go to hell." Emmanuelle sat up straight.	"Dra åt skogen!" sade Emmanuelle och satte sig kapprak.	MAP 2-1	
Smaller fry from the staff of visiting dignitaries were quartered at the Rhino... a Senegalese secretary, two men from the Ivory Coast... and there were newspapermen and a Filipino couple working for a United Nations demographic commission (Dando pointed them out) with friends from the Ghanaian Embassy.	Obetydligare medlemmar av de gästande dignitärernas staber hade inkvarterats på Silver Rhino. En senegalesisk sekreterare, två gentlemän från Elfenbenskusten. Vidare en del journalister och ett par från Filippinerna som ingick i en demografisk FN-kommission (Dando visade honom på dem) jämte några vänner från Ghanas ambassad.	MAP 1-3	
Change the settings to meet your needs. Point to the title-bar icon. Double-click.	När du är klar med ändringarna stänger du blocket genom att dubbelklicka på rubrikradsikonen.	MAP 3-1	(From IBM's OS/2 User's Guide)
However, if disk space is limited, you can install the regular typeface and ATM will approximate the other styles. Keep in mind, however, that ATM will not approximate the other styles if you are printing to a PostScript printer.	Om det är ont om plats på hårddisken väljer du bara grundteckensnittet. Varianterna blir då approximerade (gäller inte PostScript-skrivare).	MAP 2-2	
To cascade or tile all the windows whose titles appear in the Window List using a pop-up menu: Point to an empty area on the desktop. Click mouse buttons 1 and 2 at the same time.	Gör så här: Ordna samtliga öppna fönster: Ta fram Aktiva sessioner (genom att klicka med båda musknapparna på en ledig yta på skärmen).	MAP 3-3	

Length-based sentence alignment can deal with the simple mapping relations (1-1, 1-2 and 2-1) with good accuracy, but there is a problem in distinguishing between a complex relation on the one hand (n-1, n-n, 1-n) and a 1-1 relation followed by a deletion or an insertion on the other hand. The aligned text shown in Table 4 illustrates a portion of a program manual translation where a purely statistics-based aligner would have difficulties in assigning the correct links.

Table 4. Deletion and insertion relations (1-0 and 0-1)

Source	Target	Map
Scramble	**15-spel**	MAP 1-1
Scramble is a small puzzle that is solved when you arrange the puzzle pieces in the correct order.	15-spelet går ut på att lägga brickorna i rätt ordning.	MAP 1-1
You must move the pieces of the puzzle in one direction at a time.		MAP 1-0
	Välj Öppna på menyn Arkiv och sedan motiv: siffror, katter eller OS/2-logotypen.	MAP 0-1
	Välj Blanda på menyn Arkiv innan du startar.	MAP 0-1
The mouse cursor turns into an arrow when you select a piece.	Du flyttar pusselbitarna med musen.	MAP 1-1

The problem with alignment based purely on statistics is that in passages with sentences of roughly the same length, two minor perturbations can cause the alignment of the particular passage to go wrong. Many approaches to sentence alignment require that the text is synchronised on the paragraph level before the actual sentence alignment takes place. For texts with short paragraphs the risk of misalignments is small, but the longer the paragraphs are, the risk of making wrong alignments increases drastically. In hybrid approaches, this risk can be reduced by using the above mentioned anchor words or lexical cues.

In principal, all current methods for aligning texts on the sentence level are based on the following assumptions (Véronis 1999a):

1. that the two texts are ordered in the same way (or very closely)
2. that the translation contains a limited number of omissions and additions
3. that the large majority of alignments are 1-1, and that the few existing *m-n* alignments are restricted to small *m* and *n* values (≤ 2).

When a source text and target text differ too much in structure, the standard alignment methods will not produce the desired results. For example, aligning alphabetically sorted glossaries will be more or less impossible for a purely statistical alignment system. A recent approach is presented by Fluhr et al. (1999) who advocate an information retrieval solution whereby the texts are first converted into databases. The alignment task is then reduced to a "multilingual query problem" where the goal is to retrieve the target text sentence that is the best match for the "query" (i.e., the source text sentence). For texts marked-up in SGML or HTML it is possible to use the mark-up as clues for sentence alignment. This method has been applied by Martinez et al. (1998). Another approach is to use interactive alignment as proposed in Chapter 4 of this thesis. Interactive alignment will stop the alignment process at "difficult" passages of the text and ask the user to confirm the proposed alignments from the system, or make some modifications. The interactive approach is slower but will result in accurate parallel texts, which is essential for more qualitative studies of small to medium-sized parallel corpora.

In the recent ARCADE project, twelve different sentence alignment systems were compared and evaluated (Véronis and Langlais 1999). The conclusion was that the state of the art systems within sentence alignment can produce alignments with at least 95 per cent recall and precision for source and target texts that are in related languages and compliant with the three assumptions made above.

2.6 Tools related to translation and translation corpora

Tools related to translation and translation corpora should be thought of as applicable from various points of view. One and the same tool could be used by for example a translation scholar, a language engineer, a contrastive linguist and a lexicographer for their own different purposes. The basic resource, the translation corpus, is a common resource that all parties share an interest in, but the exact setup of the corpus in terms of what kind of information to include in such a corpus may differ depending on the viewpoint.

In the previous section, the sentence alignment tools needed to build up translation corpora have been discussed. As mentioned above, the translation corpus is the primary object in focus for all parties; however, there are three other types of tools that have will be discussed in this section. These three types of tools are:

1. Diagnostic tools
2. Data acquistion tools – bilingual concordance programs and word alignment systems
3. Evaluation and proofing tools.

In an industrial translation project, there is a need to be able to analyse and diagnose source texts before the actual translation starts. The characteristics of the source text may determine if and what translation tools should be used to minimize costs and assure quality. Questions about repetitions in the source text and similarities to archived translation memories may decide whether translation memory tools are to be used and, if so, what translation memories best fit the new source text. Information on vocabulary in the source text will also provide valuable knowledge to decisions on how much effort should be spent on updating term banks, etc. The *diagnostic tools* are important to translators and translation teams; however, information on repetitions and vocabulary profiles in the source texts can also provide interesting information for translation studies.

When translations have been created as a bitext, either as a translation corpus or a translation memory, then this bitext can be seen as a resource for translators, lexicographers, language engineers, translation scholars and contrastive linguists. By using *data acquisition tools*, such as *bilingual concordance programs* and *word alignment programs*, the translation corpus can be exploited and relations between various source and target text objects can be revealed on various levels. For example, the language engineer who is building a machine translation

application can extract bilingual lexicons using word alignment programs, lexicographers may use bilingual concordance programs and lexicon extraction tools to update bilingual lexicons and illustrate language use with examples from the translation corpus. Terminologists may want to extract technical terms and their translations and update term banks, and the translation scholar may want to test hypotheses about the relationships present in the source and target texts.

In practical translation production there is always a need for post-editing, especially in large translation projects involving several translators. Post-editing requires a great deal of effort to make sure that the quality meets the expectations of the customer. Normal proofing tools for monolingual texts, such as spell checker and grammar checkers, cannot capture mistaken translations if the target text is correct. Therefore *evaluation and proofing tools specifically designed for bitexts* could be designed where the objective is to capture mistaken terminology and inconsistencies in the translations, relative to the source text. It would also be desirable to build in the capacity to check how well a translation conforms to a given style-guide for a specific text type or company.

If we adhere to Kay's call for "modesty" and "only do what we know we can do reliably", we can expand the capacity and functionality for these types of tools step-by-step and integrate them in the overall translation process.

Let us take a closer look at the characteristics of categories of diagnostic tools, word alignment programs, bilingual concordancing programs and evaluation and proofing tools for bitexts.

2.6.1 Diagnostic tools

The use of diagnostic tools to determine text profiles is considered in three important phases of industrialised translation:

- Decision-making (what method or methods are appropriate with a given text-type, and with different parts of a given text?).
- System configuration (what data is relevant and how should it be acquired and put to use?).
- Post-editing (what effects is the chosen method likely to have on the target text?).

The overall goal for the diagnosis is to arrive at a description of the source text that is to be translated and to compare these characteristics to whatever available resources and tools the translators have at their disposal. Diagnosis can be performed on the current source text (text to be translated) and/or previous translations (including a related source text and its target text). Thus, we have the possibility to analyse monolingual or bilingual texts that will put different requirements on the tools in question.

Monolingual analysis can include frequency analysis of word forms, lemmas, recurrent sentences and phrases. With more sophisticated linguistic resources, it may involve recovering data on the distribution of syntactic constructions, etc.

Apart from mere frequency information, diagnosis could involve acquiring data on the distribution of various units. For instance, if you want to see how large texts are related in terms of word distributions and recurrent sentences and phrases, a "map" telling you about the degree of similarity between different chapters or subdocuments could be very useful for the translation co-ordinator who is responsible for dividing the source texts among different translators. It is not uncommon that an industrial translation project involves hundreds of documents, and the distribution of the source documents to individual translators can be made more appropriately if the co-ordinators are aware of the similarities between the various documents.

Analysis of a previously translated and related text compared to the current source text could be performed to get data on the similarity between the new text to be translated and the source text of a previous translation. This would provide data on for example if the old translation should be used as a resource (translation memory) for the new translation.

Isabelle et al. (1993) have in their system description a component called TransBase which is a translation database where source and target language elements are on the one hand stored separately, and on the other hand it holds a sentence level "translation map" where the correspondences between the source and target sentences are stored.

If diagnostic tools are tuned to specific MT systems, it is possible to compare how well the MT system will perform on the source text. Some commercial translation packages come with an application-specific diagnostic utility. Logos Corporation has, for example, developed a *Translatability Index*, which automatically assesses the suitability of a source text as a potential input to its MT system (Gdaniec 1994). The factors they consider in calculating the Translatability Index are, for example, sentence length, degree of syntactic complexity and discourse characteristics. It does not pinpoint specific problems with the text, but gives the general characteristics of the text in relation to the capabilities of the Logos MT system. The data received from this utility is a relative score which makes it possible to determine that document A is more suitable for machine translation than document B, for example.

Another way of diagnosing a text is to automatically decide the text type it belongs to. This can be done by using a relatively simple set of metrics (Karlgren and Cutting, 1994) with which the text is compared to a corpus that has been classified into different text genres. The metrics used here focus on stylistic issues, such as word and sentence length, occurrence of prepositions and pronouns, etc. More interesting from a translator's point of view would be to be able to produce statistics on how related a text is in terms of content to other texts in a corpus. For this purpose, a closer examination of content words would be necessary.

2.6.2 Word alignment programs

Previously (in section 2.5) different approaches to paragraph and sentence alignment were described. Such tools are indeed the corner-stone of parallel resources as they form the basis for other resources, such as bilingual lexicons and term banks as well as bilingual concordancing.

During the nineties there has been a great deal of research interest into word alignment systems that take parallel texts as a basis for creating correspondences between units below the sentence level, such as phrases, terms and words (for example, Brown et al. 1991, Kay and Röscheisen (1993), Smadja (1993), Kupiec (1993), van der Eijk (1993), Fung and Church (1994), Wu (1995), Chang and Ker (1996), Kaji and Aizone (1996), Macklovitch and Hannan (1996), Melamed (1995, 1996a, 1996b, 1996c, 1997a, 1997b, 1998b, 1999), Fung and McKeown (1997), Kitamura and Matsumoto (1996), Tiedemann (1997), Resnik and Melamed (1997), Hull (1998), Gaussier (1998), and Véronis and Langlais (1999)). In 1998 the ARCADE project set up a kind of competition where several word alignment systems were tested in the "word track" of the ARCADE project (Véronis 1998), see also section 7.2.1.

The overall objective for word alignment systems is to arrive at descriptions of how words and units of words in the source text are related to words and units of words in the target text. In most approaches, the idea is to link the actual tokens in the text, i.e., to find correspondences for as many text units as possible. An example from the Linköping Word Aligner software, LWA, (described in chapter 7) looks as follows:

Source: THIS PROBABLY MAKES THEM LESS ROWDY .

Target: DET GÖR DEM FÖRMODLIGEN MINDRE BRÅKIGA .

Links:

this <=> det	(1 => 1)
probably <=> förmodligen	(2 => 4)
makes <=> gör	(3 => 2)
them <=> dem	(4 => 3)
less <=> mindre	(5 => 5)
rowdy <=> bråkiga	(6 => 6)

Figure 5. Example of text links (word alignment) from a Swedish translation of Saul Bellow's *To Jerusalem and Back*.

The numbers to the right of the word correspondences indicate the word positions in the respective source and target sentences.

When as many tokens as possible in the translation have been linked, it is straight-forward to generate a bilingual lexicon, which can be seen as generalisation of the token links, as it contains all the different link types between the source and target texts. The lexicon produced will then be a translation-specific lexicon for the parallel text and can be used as a resource in the translation of similar, or related, translation projects.

An extract of an automatically generated bilingual lexicon is shown in Table 8 below.

Table 5. An extract of an automatically aligned lexicon from a Swedish translation of Saul Bellow's *To Jerusalem and Back*

Source item	Target item
less	'mindre'
let	'låt', 'låter', 'lät', 'låta'
lethargic	'letargiska'
lethargy-inducing	'slöhetsframkallande'
letter	'brev', 'brevet', 'insändare'
letters	'brev'
letting	'låter'
lev	'lev'
levantine	'levantinska', 'levantinskt'
levantines	'levantiner'
level	'nivå'
lewis	'lewis'

Another important step for creating resources is to extract terminology lists. As many terms may be new to a translation task and furthermore the terms may consist of more than one word, the translation process would be made easier if a) the terms could be identified in the source text and/or b) the terms could be analysed in a previous parallel text in order to produce translations of the already identified terms.

The first step, namely to identify terms, can be extended to the problem of identifying collocations and phrases in the source text. To produce such lists is not a trivial problem. Statistical methods based on frequency or by measuring mutual information scores for strings of words (cf. Choueka 1988, Smadja 1993, Nagao and Mori 1994, Johansson 1996, Kita et. al 1994, Dagan and Church 1997, Shimohata et al. 1997, Yamamoto and Church 1998, Zhou and Dapkus 1995) can produce lists of term and phrase candidates made up of multi-word units, but there is always a possibility that units with low frequencies will be left out. Furthermore, some kind of filtering is necessary, for example by eliminating function words before, inside or at the end of the candidate (cf. Merkel et al. 1994). Other approaches involve grammatical or syntactic processing and requires that the text is tagged for parts-of-speech which will make it possible to find candidates of low frequencies based on for example noun phrase patterns or other specified criteria (cf. Kupiec 1993; Chen and Chen 1994,) or designated noun-phrase extractors such as NP Tool from Lingsoft Oy (Birn 1997)).

The second step, to produce terminology and lexicon resources from previous translations stored as parallel texts, gives rise to some different approaches. The first option is to ignore pre-processing of multi-word units and to integrate this process into the word linking program. The second approach is to generate only the multi-word units in the source text and estimate their correspondences in the source text, that is, without specifying lists of possible target candidates. The third alternative is to process both the source target text in similar ways and

feed the word linking program with lists of multi-word units for both the source text and the target text. The first approach has been adopted by Melamed (1997a) and produces impressive results, but is inefficient in terms of processing time; the second approach was adopted by for example Smadja (1993) and is more efficient, but lacks the flexibility of Melamed's approach. The advantage of the third approach is better efficiency but less flexibility; the successful linking of multi-word units depends highly on the quality and size of the pre-processed term and phrase candidates.

For industrial translation projects, such as the translation of technical manuals, terminology lists of reasonable quality can be useful as a stage in creating and updating terminology banks for customers or in-house translation projects. To have information of what the potential terms look like, their frequency and distribution in various documents can provide the terminologist with insights that makes a semi-automatic approach to building term banks far more useful than searching manually for potential terms that should go into the term bank. Even if the translation co-ordinator or terminologist only has access to potential terms and collocations in the source text, this could prove to be a useful starting point for deciding what actually should go into the term bank. If a preliminary list of candidates is reviewed at the same time as the terminologist has access to a bilingual concordancing program (such as Termight (Dagan and Church 1997) or TransSearch (Macklovitch 1992), the possibility of producing high-quality term banks increases drastically.

2.6.3 Bilingual concordance programs

Earlier bilingual concordance programs were mentioned as a tool that can be used during the creation of specific lexicons and term banks. Bilingual concordance programs can indeed be used at that stage, but are perhaps more appropriate as tools for acquiring data from previous translation during the actual translation phase. Perhaps the best known bilingual concordance program is TransSearch from Montreal, illustrated in Figure 6.

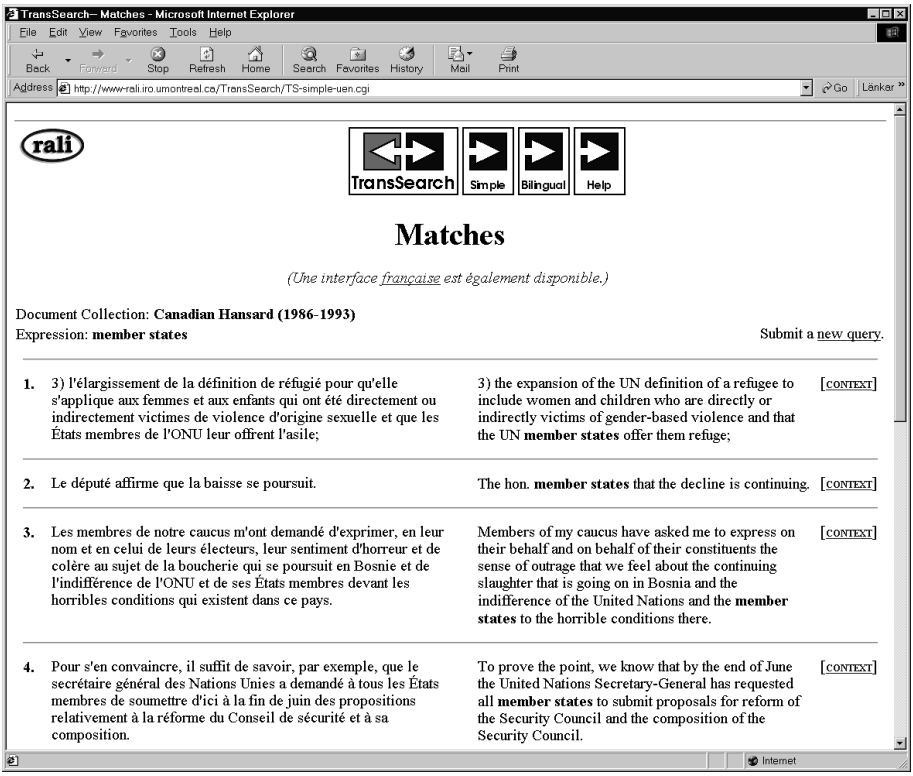


Figure 6. TransSearch from RALI, Montreal

Figure 6 illustrates a search for the English phrase "member states" and shows the phrase in context with the corresponding French sentence to the left.

In Sweden, a similar concordance tool has been made available by Gothenburg University within the Pedant project (Ridings 1998). An example of a search with the Pedant concordance tool is shown in Figure 7. Here the Swedish word "hindra" is searched for in the parallel text and by looking at the English corresponding sentences, the user could discover that "hindra" corresponds to "block", "hinder", "restriction" or "restrict" in the first four sentences.

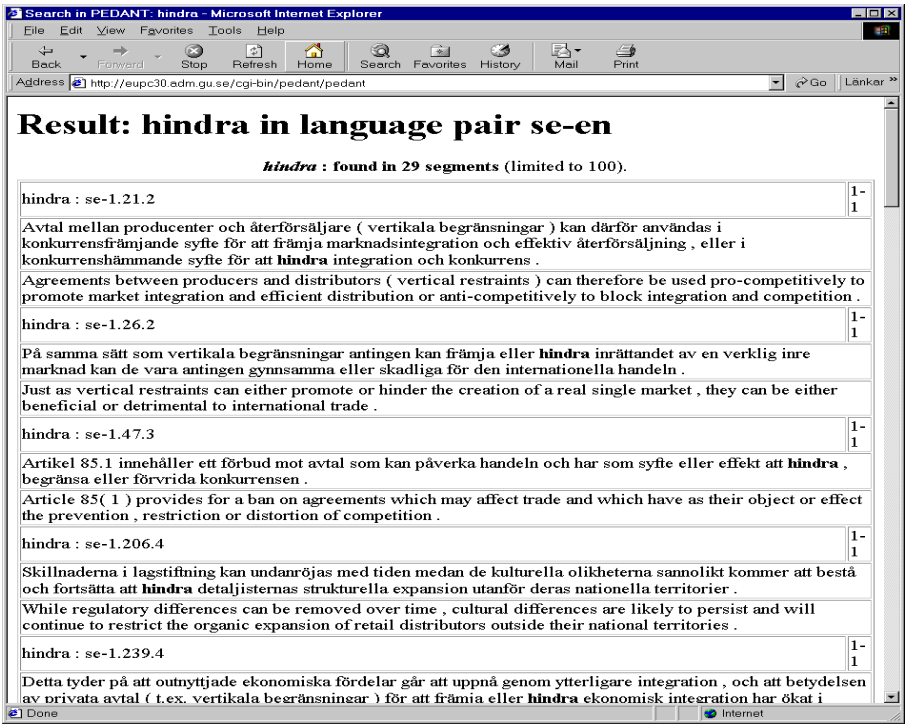


Figure 7. Pedant Bilingual Concordance program

An example of non-web-based, stand-alone bilingual concordance program is ParaConc from Athelstan (Barlow 1995). Other types of data acquisition tools are Translation memory workbenches (for example, Trados Translator's Workbench, IBM's Translation Manager and others), as well as Terminology lookup systems (such as Trados Multi-Term) where users can import their own resources and use the data acquisition tools as an integrated part of their word processing program. The parallel texts used and incrementally built up when using translation memories can in most cases also be browsed through in the same manner as the concordance tools described earlier.

2.6.4 Evaluation and proofing tools for bitexts

During the post-editing phase, when translations are to be proofread and validated, tools that could enhance and shorten this often tedious and time-consuming phase would be welcomed by many people working with translations. Most larger translation customers, translation companies and publishing houses issue their own translation guidelines and require that their translators adhere to these guidelines (Microsoft 1993, Ström and Windfeldt 1991). Often these guidelines contain long lists of "do's" and "don'ts" but the overall aim is to encourage the translators to use consistent terminology, consistent style and to

minimise the number of errors in the translations. The major difference between proofing an original text compared to proofing a translation is that in the first case it is sufficient to scrutinise "one" text in one language, whereas proofing a translation, at least in the ideal world, involves not only checking language-related problems within the translation, but also ensuring that the translation represents the content expressed in the source text. This means that proofing translations requires ways of having access to both the source text and the target text at the same time. At least one translation company I am aware of solved this by duplicating all the paragraphs in the source document before it was released to the translators. Then the translators were asked to translate only the first of each pair of identical paragraphs. (The second paragraph could also be hidden on the screen, and then during the proofreading, the proofreaders had access to both the original and the source text on the same printed document or on the screen.)

Automating the evaluation and proofing process of translated documents is therefore a more complex problem than automating "normal" proofreading. However, there are approaches that could improve this stage.

First, and perhaps a little too obvious, spelling correction and available grammar checkers could assist in pinpointing some of the mistakes made in the target text. Secondly, relatively simple tools that could check the translations for false friends and inconsistencies could be applied. At RALI in Montreal a program called TransCheck can identify relatively simple common mistakes that translators make when translating between French and English (Macklovitch 1994, 1995). Within the near future, I expect that we will see a number of tools that could identify terminological errors, given a validated term bank, and verify that certain required stylistic features are present in the target text.

2.7 Summary

In this chapter I have reviewed recent developments in the fields of translation studies, corpus linguistics and computational linguistics pointing out relationships that make all the fields profit from each other.

In order to produce better translations, we must know what characterises high-quality translations. By studying what professional translators produce and how the source text is related to the original, we can begin to understand what is needed in terms of translation support tools and, perhaps also, automatic translation systems. This is indeed an empirical field of science and the resources now made available to us in the form of huge quantities of text in electronic format together with powerful hardware must be put to good use. A translation program can produce grammatical translations that are deemed to be useless by translators or potential customers. The question then is why the output from such systems fail to satisfy the users. If we can answer this, we have the platform to start to build something else. Corpus linguistics provides understanding in how to set up and analyse large quantities of corpus data. Translation studies provides theoretical insights into the nature of translation and the taxonomies to describe the relationship between source and target texts. Computational

linguistics provides the cornerstones for how to implement the tools that we need in efficient ways. However, it is only by putting all these fields together and making them interact that we can gain knowledge about how the new generation of translation support tools should be designed.

3 Consistency and the translator

In the translation industry, the use of translation support tools, such as translation memories, have been increasing steadily in recent years. Translation memories are a straight-forward implementation of the idea of *reusability* (see chapter 1). As mentioned in the previous chapter, Translator's Workbenches, to which category translation memory systems belong, are a type of application advocated by computational linguists as an alternative to fully automatic translation. In this chapter, an empirical study of how human translators react to consistency and to the notion of reusability in technical translations is presented. Furthermore, the translators' attitudes towards translation tools are discussed.[4]

When large quantities of technical texts are being translated manually, it is very difficult to produce consistent translations of recurrent stretches of text such as paragraphs, sentences and phrases. This can have different causes, for example, several translators work on different sections of the same document simultaneously, the source text is not final and may be changed at a later stage, it may be too time-consuming or practically impossible to manually identify recurrent units in the source text. Individual translators making up a translation team will also have individual criteria for choosing a certain translation or even choosing from a set of possible translations.

One suggested remedy to the problem of consistency in translation is to use tools based on translation memories. Successful use of such tools entails the supposition that source repetitions can be transferred to the target text. To validate this supposition, variant translations in two software manuals were identified and categorised. Then a questionnaire based on the material was designed and distributed to the translators at the two translation companies which had translated the manuals. Apart from the translators, we also involved the project leaders at the translation companies and a representative from the customer (the software company).

Translation memory tools are used to the best advantage when the following conditions hold:

1. The source text is highly repetitive internally
2. The source text is a new version of a text that has already been translated and exists as a translation memory (i.e., external repetition).
3. Repeated sentences (or segments) of source text are in principle transferable to corresponding repeated sentences in the target text.

[4] This chapter is an extension of the work presented in Merkel (1998).

The first two conditions can quite easily be checked automatically, which was done in a previous study (Merkel 1992). By analyzing two software manuals for repetitions it was concluded that both internal repetitions and external repetitions were of considerable proportions and that it indeed would be valuable to reuse translations, in a manner similar to what translation memories can do. The study of two versions of a spreadsheet program manual revealed an internal repetition rate of up to 25 per cent (i.e., 25 per cent of the sentences in one of the manuals were among the set of repeated sentences). See Table 6 below.

Table 6. Internal repetition rates for two spreadsheet program manuals (SS1 and SS2)

Internal Repetition	**SS1**	**SS2**
Total no of words	254,350	248,089
Different word types	5,362	5,691
Internal Sentence Repetition Rate	25 % (2,290 sents)	15 % (1,822 sents)

When the two manuals were compared with each other (for external repetitions), the study revealed that 20 per cent of the sentences were identical in the two texts (see Table 7 below).

Table 7. External Recurrence in SS1 relative to SS2

External Repetitions in SS1 relative to SS2	
Shared sentences	3,677
External Sentence Repetition Rate	20 %

If we make the hypothetical assumption that a translation memory tool had been used for translating the second version of the manual, it is obvious that considerable efficiency gains could have been achieved. But, note that we have to be sure that when such a tool is used, it rests on the third condition mentioned above, namely that repetitions can indeed be transferred from the source to the target text. This third condition is harder to verify, but in this chapter we attempt to do this by confronting translators with their own translations and by asking them how they value repetitions in the source and target texts.

The study should be seen as complementary to assessments of practical usage of computer-assisted translation tools (Schäler 1994) and Vasconcellos (1994). To prepare the text material, the DAVE package (which will be described in Chapter 4) was used for detecting recurrent sentences and phrases as well as for aligning the software manuals on the sentence level. At the time of the study the discrepancy module included in DAVE was not ready (a module that pinpoints inconsistent translations) which means that the variant translations used in the study were detected by manually scanning sorted translation databases containing recurrent sentences. The discrepancy module present in DAVE is similar to TransCheck (Macklovitch 1994) in that it operates on bitexts (i.e.,

representations of translations with explicit links between source and target sentences), but the module operates on consistent or inconsistent translations of sentences rather than lexical items.

3.1 Objectives and methodology

The study in this chapter focuses on the distribution of translations of recurrent source sentences and how translators familiar with the text type evaluate the existing translations of the recurrent units in different contexts. To collect the necessary data, I designed a questionnaire with two parts. The first part included questions about translation support tools in order to provide important insights into the usability of translation memory tools or techniques, and the respondents' attitudes towards the benefit of using such tools. The second part consisted of 50 examples from two computer manuals from the same company, where each example described an identical source segment (paragraph, sentence, heading, etc.) occurring in two different contexts. These source segments were shown with the corresponding target segment, which had been translated in two different ways. The task for respondents was to decide whether they would prefer consistent translation in the two contexts or the two variant translations. If consistent translations were preferred, the respondent was asked to motivate the choice and also rank the different alternatives. If variant translations were preferred, the respondent was asked to justify this choice. The questions accompanying each example were identical and looked as follows:

1. Would you prefer a consistent Swedish translation of the English example in both A and B?	
❑ Yes	❑ No (if you answer No, go to question 3.)
2. How would you rank the translations? Answer this only if you have chosen **Yes** on question 1.) ❑ They are equivalent. Both A and B are good translations. ❑ I prefer A to B. ❑ I prefer B to A. ❑ I prefer another consistent translation. Justify your choice: _____ _____ _____ _____	3. Justify why you want different translations. (Answer this only if you have chosen **No** on question 1.) _____ _____ _____ _____ _____ _____ _____ _____

Figure 8 Questions accompanying each example in part 2 of the questionnaire

To avoid any influence caused by the way the examples were ordered, the order of the 50 examples was randomized for each respondent.

The questionnaire was sent out to both in-house and freelance translators working for two Swedish translation companies. One requirement was that translators should be familiar with the text type and the requirements of the translations (i.e. that they should have translated texts from the particular software company before). One project manager at each of the translation companies was also asked to complete the questionnaire as well as the person responsible for translation quality at the client software company. In total 13 completed questionnaires from translators were answered and returned (8 from freelance translators and 5 from in-house translators). We also received feedback from the two project managers and the client software company.

To distinguish between different categories of translation variants, the examples used in part two of the questionnaire were divided equally (five examples from each category) over the 50 examples, as shown in Table 8. I distinguish between three major types of translation variation:

A. **synonym variants** (where the variants have the same underlying logical form),

B. **partially synonym variants** (where the variants differ in the degree of specification) and

C. **non-synonym variants** (where the variants do not have the same underlying logical form, i.e. there is conflicting semantic content).

A fourth category was also included where longer stretches of text were repeated (category D below). The numbers in brackets indicate the example numbers used in Table 19 and Table 22.

Table 8. Variant translation categories

Major type of variation	Variation categories	Examples
A. Synonym variants	1a. Syntactic variants – same context types (No. 1-5)	Running text-running text, heading-heading, etc.
	1b. Syntactic variants - different context types (6-10)	Running text–heading, table cell–heading, etc.
	2a. Morphematic variants – same context types (11-15)	Running text-running text, heading-heading, etc.
	2b. Morphematic variants – different context types (16-20)	Running text-heading, etc.
	3. Lexical variants (21-25)	Non-terminological synonyms
	4. Coherence variants (26-30)	Pronouns, adverbs, etc.
B. Partially synonym variants	5. Specification variants (31-35)	Less or more specific content
C. Non-synonym variants	6. Terminological variants (36-40)	Varying terminology
	7. Content variants (41-45)	Erroneous content translations
D. Recurrent multi-sentential segments	8. More than one recurring subsequent sentence (46-50)	Repeated paragraphs or whole sections.

A. Synonym variants

The synonym variants categorized in four different groups (1-4) depending on whether the synonymity stemmed from syntactic choices, morphematic variation, use of lexical synonyms or the use of coherence markers such as pronouns and adverb. As the text material included a certain type of syntactic and morphematic variation depending on structural context, groups 1 and 2 were divided into two subgroups where the *a* subgroups contained variants in the same type of context (for example, if both variants were used as headings, or

both were used inside instructions). The *b* subgroups then contained variants where the structural contexts differed, for example if one variant was used in a heading and the other in a cell table.

The examples below have been shortened compared to the questionnaire that was distributed, but the respondents were presented with the following type of text data where the repeated source sentence and the corresponding variant translations were underlined.

In Table 9 the variation in translation choice is one of word order. The A translation variant contains the object "frågeverktyget" in the initial position, whereas the B variant has the subject "Du" in the initial position. In this case the variants occur in exactly the same type of structural context, namely inside a caution.

Table 9. Syntactic variants - same context types (1a)

Variant	Source	Target
A	**Note** The Query Builder is available for working with queries only.	**Obs!** Frågeverktyget kan du bara använda när du arbetar med frågor.
B	**Note** The Query Builder is available for working with queries only.	**Obs!** Du kan bara använda Frågeverktyget när du arbetar med frågor.

An example of syntactic variants in different contexts is shown in Table 10. Here the A variant functions as a heading whereas the B variant occurs inside a table cell.

Table 10. Syntactic variants - different context types (1b)

Variant	Source		Target	
A	**Calculating with precision as displayed:** • Affects all worksheets in the active workbook. • Does not affect numbers in the General format, which are always calculated with full precision.		**Vid beräkning med visad precision:** • Påverkas alla kalkylblad i den aktiva arbetsboken. • Påverkas inte tal i formatet Standard eftersom dessa alltid beräknas med exakt precision.	
B	Type this keyword and choose Show Topics	Select a topic and choose Go To	**Skriv detta nyckelord och välj Visa avsnitt**	**Markera ett avsnitt och välj Gå till**
	calculating formulas	Calculating a portion of a formula	beräkna formler	Beräkna delar av formler
	calculating with displayed precision	Calculating with precision as displayed	beräkna med visad precision	Beräkna med visad precision

Morphematically variant translations are translations where the morphological form of one or several words are different. In Table 11 the difference between the A and B translations is the number of the noun "nivå" (Eng. 'level').

Table 11. Morphematic variants - same context type (2a)

Variant	Source	Target
A	Removes tracer arrows from one level of precedents. Subsequent clicks remove the next level of arrows.	Spårningspilar för en nivå av överordnade celler tas bort. Om du fortsätter att klicka på den här knappen tas pilarna för efterföljande nivå bort.
B	Removes tracer arrows from one level of dependents. Subsequent clicks remove the next level of arrows.	Spårningspilar för en nivå av underordnade tas bort. Om du fortsätter att klicka på den här knappen tas pilarna för efterföljande nivåer bort.

Another example of morhematic variants can be seen in Table 12 below. Here the term "inbäddat diagram" is in the singular in variant A (heading), and in the plural in variant B (table cell).

Table 12. Morphematic variants - different context types (2b)

Variant	Source		Target	
A	**Moving and Sizing an Embedded Chart** To move the chart on the worksheet, select it by clicking anywhere in the chart and then drag it where you want it. If the chart is active, you can move it by dragging its border.		**Flytta och ändra storlek på ett inbäddat diagram** Om du vill flytta diagrammet i kalkylbladet markerar du det genom att klicka i diagrammet och sedan dra det till en ny position.	
B	**Type this keyword and choose Show Topics**	**Select a topic and choose Go To**	**Skriv detta nyckelord och välj Visa avsnitt**	**Markera ett avsnitt och välj Gå till**
	charts, embedded	Moving and sizing an embedded chart	diagram, inbäddade	Flytta och ändra storlek på inbäddade diagram
	sizing chart sheets	**Sizing a chart sheet with the window**	ändra storlek på diagramblad	Ändra storlek på diagramblad med hjälp av fönstret

As the material was in the technical domain, it is essential to distinguish between terminology which is dependent on the domain and lexical items which belongs to general usage of the language. Therefore, a distinction was made between use of synonyms which are non-terminological (group 3) and variation of terminology (group 6). An example of non-terminological variation is shown in Table 13 below. The original word "enter" is translated with "anger" in A and with "skriver" in B.[5]

[5] In program manuals such as the ones used in this study, it is common to regard system-specific noun phrases as belonging to terminology as well as all text strings appearing in the program (such as names of commands, menus, buttons, etc.). The criterion is that they refer to one "object" or "action" and that there is no alternative option. Although the English "enter" is used in a technical sense, the source writer could just as well have chosen "type" to get the same meaning across.

Table 13. Lexical variants - non-terminological synonyms (3)

Variant	Source	Target
A	From the Tools menu, choose Scenarios, and click the Add button. Type a name for the scenario (such as "Best Case" or "Mary's Assumptions"). In the Changing Cells box, enter the references or the names for the cells that you want to change. If you enter more than one reference, separate each reference with a comma.	Välj **Scenarier** på **Verktyg**-menyn och välj Lägg till. Skriv ett namn för scenariot (t ex Bästa fallet eller Marias antaganden). I rutan "Justerbara celler" anger du referenser eller namn för de celler som du vill ändra. Om du anger mer än en referens måste du använda semikolon mellan referenserna.
B	From the Tools menu, choose Scenarios, and then choose the Summary button. The Scenario Summary dialog box appears. (The Summary button is unavailable if there are no scenarios defined for the current worksheet.) In the Result Cells box, enter the references or the names for the result cells that you want to appear in the report. If you enter more than one reference, separate each reference with a comma.	Välj **Scenarier** på **Verktyg**-menyn och välj sedan Sammanfattning. Dialogrutan **Sammanfattningsrapport** visas. (Knappen Sammanfattning är inte tillgänglig om inga scenarier har definierats för det aktuella kalkylbladet.) I rutan "Resultatceller" skriver du referenser till eller namn för de resultatceller som du vill visa i rapporten. Om du skriver mer än en referens måste du skriva semikolon mellan referenserna.

The fourth synonym category has to do with the presence of coherence markers in the translations. In Table 14 the only difference between variant A and B is that A contains the adverbial "nu" (Eng. 'now'), which does not change the underlying logical form of the translation as it could be said to be implicit in the context of B.

Table 14. Coherence markers (4)

Variant	Source	Target
A	2. Select Form And Report Wizards from the list of add-ins. 3. Choose the Customize button. Microsoft Access displays the Customize Add-in dialog box.	2. Välj Formulär- och rapportguider på listan över tillägg. 3. Välj Anpassa. Nu visas dialogrutan **Anpassa tillägg**.
B	2. Select Form And Report Wizards from the list of add-ins. 3. Choose the Customize button. Microsoft Access displays the Customize Add-in dialog box.	2. Välj guiden Formulär och rapporter i listan "Tillgängliga bibliotek". 3. Välj Anpassa. Dialogrutan **Anpassa tillägg** visas.

B. Partially synonym variants

The second main category contains variants where the underlying logical forms of the two variants are not exactly identical, instead they are partially overlapping. This overlapping is usually expressed as a difference in the degree to which an object or relationship is specified. Typically, one of the variants is either more specific or less specific than the other. In Table 15 there is a difference in the degree of specification between "samtliga kontroller i gruppen" (A) and "samtliga markerade kontroller" (B).

Table 15. Degree of specification (5)

Variant	Source	Target
A	In the property sheet, change the appropriate properties. Microsoft Access changes the property settings for all the selected controls.	Ändra egenskaperna i egenskapsfönstret. Egenskapernas inställningar ändras för samtliga kontroller i gruppen.
B	In the property sheet, change the appropriate properties. Microsoft Access changes the property settings for all the selected controls.	Ändra egenskaperna i egenskapsfönstret. Egenskapernas inställningar ändras för samtliga markerade kontroller.

C. Non-synonym variants

The third main category has to do with variations that are not synonymous, i.e., there is a difference in the underlying logical form. Terms are the objects and relationships that have one single linguistic representation, such as names of objects and also names of parts of the documentation. In Table 16 the name of

chapter 13 is translated differently in the two variants, which is classifed as a terminological variant (only one of the two can be correct).

Table 16. Terminological variants (6)

Variant	Source	Target
A	You can add text boxes, arrows, and other graphic objects to charts. For more information, see Chapter 13, "Creating Graphic Objects on Worksheets and Charts."	Du kan lägga till textrutor, pilar och andra grafiska element i diagram. Mer information finns i kapitel 13, "Skapa grafiska objekt i kalkylblad och diagram".
B	The Object and Placement commands on the Format menu are available to use with embedded objects. With these commands, you can control properties, including graphic patterns, protection, and positioning relative to cells. For more information, see Chapter 13, "Creating Graphic Objects on Worksheets and Charts."	Kommandona **Objekt** och **Placering** på **Format**-menyn är tillgängliga när du arbetar med inbäddade objekt. Med dessa kommandon kan du styra egenskaper, bl a grafikmönster, skydd samt placering i relation till celler. Mer information finns i kapitel 13, "Grafiska objekt i kalkylblad och diagram".

Another case of non-synonym variation has to do with the choice of contradicting lexical items or structural choices that produce different meanings. As in the previous case, only one of the two variants can be said to be a correct translation. In Table 17 below it is obvious that only one of the variants can be correct as they have complete opposite meanings due to the negation in variant A.

Table 17 . Content variants (7)

Variant	Source	Target
A	**To create an option group without a Wizard** 1. From the View menu, choose Control Wizards (if the Control Wizards command displays a check mark).	**Skapa en gruppruta utan guide** 1.Välj **Kontrollguider** på **Visa**-menyn (om kommandot inte är förbockat).
B	**To create an option group without a Wizard** 1. From the View menu, choose Control Wizards (if the Control Wizards command displays a check mark).	**Skapa en gruppruta utan guide** 1. Välj Kontrollguider på Visa-menyn (om Kontrollguider är förbockat).

D. Recurrent multi-sentential segments

The final category that was included in the questionnaire contained a longer passage of text than a sentence (at least one paragraph) with two variant translations. In the questionnaire the respondents were given a larger context than that shown in Table 18, both before and after the paragraph in focus.

Table 18 . More than one recurring subsequent sentence

Variant	Source	Target
A	A page break control is like any other control. You can select it by clicking it. After you select it, you can copy it, delete it, or display its property sheet.	En sidbrytningskontroll är som andra kontroller. Du markerar den genom att klicka på den. När du har markerat den kan du kopiera den, ta bort den eller visa dess egenskapsfönster.
B	A page break control is like any other control. You can select it by clicking it. After you select it, you can copy it, delete it, or display its property sheet.	Kontrollen för sidbrytning är som alla andra kontroller. Du kan markera den genom att klicka på den. När du har markerat den kan du kopiera den, ta bort den eller visa dess egenskapsfönster.

3.2 Attitudes towards translation tools

Nine questions were included in part 1 of the questionnaire. The questions involved the respondents' general attitude towards consistency and variation, translation tools, tools to identify and translate recurrent paragraphs and sentences and recurrent fuzzy patterns (such as *From the **X** menu, choose **Y**.*) They were also asked if they used the inbuilt search-and-replace function of word processors while translating. Finally they were asked about their opinions regarding translation memory tools, a hypothetical translation verification tool that highlighted variant translations and if they had any other comments on translation.

3.2.1 The translators

All translators agreed that terminology consistency is important. Twelve out of thirteen also stated that sentences and phrases should be translated consistently and that the source text often shows too much variation. Six translators found that it was difficult to know what was actually recurrent in the source text, especially across chapter boundaries. Two translators wanted variation of recurring sentences in running (descriptive texts), but not in instructions. Still, the majority of the translators did not see variation as a goal in itself. One translator expressed this as "Variation is only confusing. The text is probably not read from cover to cover, instead the reader looks up different things."

All translators use term lists in electronic format and all of them use the printed guidelines issued by the software company. Twelve out of thirteen translators use term/word lists in printed form as well, but only two use look-up facilities directly from the word processors. None of the translators uses translation memory tools for these particular texts, but two translators used such tools for other customers' translations.

Two translators claim that they use tools to identify and translate recurrent paragraphs. Five translators use the search-and-replace function in the word processor manually to do this. The search-and-replace function is used when the

translator notices that certain segments are repeated during translation. Six translators do not use tools for this purpose.

No translators have access to tools to identify close matches (fuzzy matches) automatically. Two translators state that they use search-and-replace with wildcards occasionally.

Eleven translators use the built-in search-and-replace function of their word processor regularly; however, many of them added that they do this "with caution," or that they "always check replacement case by case". Two translators state they do not use the feature at all.

Nine translators have a positive attitude towards the use of translation memories; four show some degree of hesitation. The reservations made are that translation memories "must be easy to use", that they should be used "only as a reference", and that translators "have to be careful because terminology may have changed". One translator states the fear that it could be "tempting to work too quickly which will lead to increased number of mistakes". No translator was totally against the use of translation memories.

Nine translators were positive towards a verification tool, three had some doubts and one translator saw no use of this kind of tool. One translator noted that this would be especially useful to check terminology and expressions.

Typical comments from the translators involved concerns about the quality of the source text. There is also a fear that translation work will become more tedious and boring, that some of the creative aspects of the job will disappear with the increasing use of translation memory tools. Some translators expressed concern over the changing role of the translator, from a linguistic innovator to a linguistic operator, or as one of the respondents expressed it: "The translator is reduced to somebody who presses the OK button."

3.2.2 The translation companies

The questionnaire directed to project leaders at the two translation companies was similar to the translators' questionnaire part 1, but geared towards what the project leader wanted the translators involved in the project to do.

In translation company A, the project leader is positive towards translation tools. At translation company A, they do not use tools for handling recurrent paragraphs or sentences in general, but the use of search-and-replace functions is encouraged, if it is done with judgment. The project leader thinks that the use of translation memories would benefit the company, and the same applies to the use of a verification tool. She states that tools would make translations of, for example references to chapter and section headings easier in printed documentation and in on-line help texts. However, she concludes, when running text is involved, there is a risk that the creativity and ability to localize text can be inhibited.

In translation company B, the project leader is, in general, very positive towards translation tools. Specifically, he would like the translators to use search-and-

replace functions. At translation company B they have created their own tools for handling recurrent sentences, as well as their own terminology tools. He also thinks that the translation memories and verification tools would be used and found useful by translators and editors. However, he emphasizes that it is important that translation tools be very flexible and easy to use. Furthermore they must include functions for handling updates of the source text and provide room for a review and verification phase.

3.2.3 The customer

In an interview with the person responsible for the Swedish translations (terminology and quality assurance) at the customer company the following views were expressed:

Translation memory tools are an excellent help for repetitive texts. The prerequisite is that the previous translation memories are correct, which is not always the case. Terms may have been changed and there may be mistakes in the old translations. In general it is not advisable to reuse an old translation without verifying the accuracy of the translation. This verification must be done before a translation memory can be used in a new project. It is important that translators be able to view the context in which a certain sentence is to be translated. A pure sentence-by-sentence translation is not advisable at all. Automatic translation without the confirmation of the translator is not acceptable.

From the questionnaire, it was clear that the representative for the customer was less inclined to prefer consistent translations than the translators. In the interview, this opinion was explained by the fact that the objectives may be different for a customer and a translator. The customer demands a high quality translation and is not really interested in how this is achieved, whereas the translator is by definition involved in the actual process of translation and strives to minimize the effort of reaching high quality translation. This may result in the discrepancy shown in the questionnaire, i.e., that the customer representative had a much higher proportion of "doesn't matter" replies than the translators in general. For example, for running text (descriptive text) the customer representative stated that different translations were possible, but not necessary, whereas the translators preferred consistent translations to a higher extent.

3.3 Consistency vs. Variation

As described earlier, the respondents were asked to state whether they preferred consistent translations of a given source segment in two different contexts. The options given were either *yes* or *no*, with a space for the respondents' own motivations for his/her choice. When I examined the questionnaires, it became apparent that there was a need for a third response, in between yes and no, namely a response which we can call "doesn't matter". This applies when the translator in the justification for the choice has indicated that the translation on the one hand could be consistent, but that it would not matter whether the source segment also was translated differently or vice versa. This introduction of

a third response category may seem to diminish the possibilities of interpreting the results, but if we still regard the responses as binary (*yes* or *no*), with the judgment of consistent translations as the primary problem, the responses "yes" and "doesn't matter" both indicate the possibility of consistent translations, whereas a "no" response rules out consistency. Table 19 summarizes the results from the questionnaire regarding the respondents' choice of whether a source sentence should be translated consistently or not. The "YES" option shows how many respondents judge that the source sentence should be translated consistently. The "NO" option shows how many respondents want different translations in the two contexts and the "D M" (Doesn't Matter) option states how many respondents regard it possible, but not necessary, to be consistent.

Table 19. Responses to questions on consistency

Ex. No.	YES	D M	NO	Ex. No.	YES	D M	NO
1.	**14**	2	0	**26.**	**14**	2	0
2.	**13**	2	1	**27.**	**16**	0	0
3.	**14**	2	0	**28.**	**12**	4	0
4.	**15**	1	0	**29.**	2	6	**8**
5.	**12**	4	0	**30.**	**15**	1	0
6.	**16**	0	0	**31.**	**15**	1	0
7.	6	2	**8**	**32.**	**15**	1	0
8.	**10**	1	5	**33.**	**15**	1	0
9.	**12**	0	3	**34.**	**12**	2	2
10.	**14**	1	0	**35.**	14	2	0
11.	**14**	2	0	**36.**	**16**	0	0
12.	**12**	4	0	**37.**	**16**	0	0
13.	**14**	2	0	**38.**	**16**	0	0
14.	**14**	2	0	**39.**	**15**	1	0
15.	**11**	5	0	**40.**	**15**	0	0
16.	**10**	0	6	**41.**	**16**	0	0
17.	**10**	0	6	**42.**	**12**	3	1
18.	**11**	1	4	**43.**	**15**	1	0
19.	**11**	1	4	**44.**	**16**	0	0
20.	**11**	0	5	**45.**	**15**	1	0
21.	**13**	3	0	**46.**	**16**	0	0
22.	**14**	2	0	**47.**	**15**	1	0
23.	**13**	3	0	**48.**	**13**	3	0
24.	**14**	2	0	**49.**	**13**	3	0
25.	**15**	1	0	**50.**	**13**	3	0

A first glance at the answers indicates that there is a clear preference for consistent translations; for 48 examples a majority of the respondents prefer consistent translations. Furthermore, in 38 of the examples all respondents have ruled out variation (zero value for NO).

In two of the examples, a majority of the respondents are in favour of variant translations (example 7 and 29). The first is an example of where *Calculating with precision as displayed:* occurs as both a heading and as cell item; that is, in different contexts. The second (example 29) depicts the source sentence *This value is derived using the formula:*, where the first translation is spelled out with a definite description and the second has kept the coherence marker (*this value*) of the

source sentence (*Pearson-korrelation beräknas med följande formel*" vs. "*Detta värde beräknas med följande formel*").

To compile the found preferences related to categories, the number of responses in each category and distinguished two major classes have been added: Categories with a clear preference for consistency and categories where the respondents have shown a marked degree of hesitation towards consistency. Table 20 and Table 21 show the number of responses and the number of examples over which the response is distributed.

Table 20. Eight categories with a preference for consistency

Category	YES	Doesn't Matter	NO
1a. Syntactic variants - same contexts	*68/5*	*11/5*	*1/1*
2a. Morphematic variants -same contexts	*65/5*	*15/5*	*0/0*
3. Lexical variants	*69/5*	*11/5*	*0/0*
4. Coherence markers	*59/5*	*13/4*	*8/1*
5. Degree of specification	*71/5*	*7/5*	*2/1*
6. Terminological variants	*79/5*	*1/1*	*0/0*
7. Content variants	*74/5*	*5/3*	*1/1*
8. Multi-sentential variants	*70/5*	*10/4*	*0/0*

In category 4, there is a single example (number 29) where the respondents actually preferred variant translations, but this seems to be an exception as there was a clear preference for consistency in the other four examples. (Example 29 is discussed above.)

Table 21. Two categories with a marked degree of hesitation for consistency

Category	YES	Doesn't Matter	NO
1b. Syntactic variants - different contexts	*60/5*	*4/3*	*16/3*
2b. Morphematic variants - different contexts	*58/5*	*2/2*	*20/4*

In Table 21, the figures for the two categories where the functional contexts differed are shown. This contextual parameter seems to be the deciding parameter. There is still a majority of YES answers, but the number of NO answers is considerably higher for these categories. The trend for the categories with different functional contexts is definitely more towards the "hesitation" side than in other categories.

The responses here can only be interpreted in one way, namely, that consistency is something that technical translators aim for in general. The exception to this rule regards context. When two source sentences (or segments) occur in different structural contexts, such as headings and table cells, translators should be more cautious in applying consistent translations. These recurring source sentences may often require different target sentences. In some translation software this is indeed handled by recognizing that they have different

formatting tags (styles or other mark-up properties) which means that no perfect match will occur, only a fuzzy match which leaves the translator to make the choice of whether or not to be consistent.

3.3.1 Uniformity in deciding preferred translation alternative

In the previous sections we have seen that there is an overwhelming general tendency towards consistency. One side-effect of the questionnaire was that we received a great deal of data on how uniform the choice of alternative was between different translators. As described earlier, the translators were not only asked to state whether they preferred consistency or variation in the translation pair given, but also to state which of the alternatives they preferred. They were also asked to say whether the alternatives were equally appropriate or whether they preferred another translation. Given each example, the translator had four choices: (a) *Prefer A*, (b) *Prefer B*, (c) *Prefer A or B (equally appropriate)*, or (d) *Prefer other translation* (see Figure 8).

As the respondents were only presented with a limited context (around two paragraphs), we expected a certain degree of disagreement in the respondents' answers. However, the differences were far greater than anticipated. Table 22 below summarizes the choices from the questionnaires. A indicates the number of respondents who choose alternative A for each example, B alternative B, "Other" indicates "other translation" and EQ suggests that both A and B are equally appropriate translations. For each category, the alternative with the highest number is indicated in bold face. Note that the A and B alternatives do not signal any "degree of appropriateness" on our part, they are chosen completely randomly.

Table 22. Preferred translation alternatives

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	1	**10**	0	1	**6**	**6**	1	3	**8**	3	**8**	4	4	0	2	2	0	2	2	**5**	**8**	1	**6**	**8**	3
B	7	1	11	**10**	5	4	3	**4**	4	**6**	2	**8**	4	**12**	6	**3**	4	4	**5**	2	0	5	1	1	**6**
Other	2	2	1	4	1	5	1	2	0	0	3	0	0	1	0	1	1	1	0	1	3	2	7	1	1
EQ	**6**	2	4	1	4	1	**4**	2	1	5	2	4	**8**	3	**8**	3	**5**	**5**	4	3	5	**8**	2	6	6
No.	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
A	**11**	1	2	3	4	**9**	0	2	3	**5**	**7**	**9**	**10**	3	0	1	2	4	**14**	2	1	5	4	4	1
B	2	**12**	**7**	5	**6**	1	**15**	**11**	**7**	**5**	0	2	0	**8**	**7**	**11**	3	**7**	1	**10**	**7**	**7**	0	1	**9**
Other	1	0	1	0	0	3	1	2	0	3	5	1	1	1	5	4	**6**	5	0	1	2	2	**7**	2	3
EQ	2	3	6	0	6	3	0	1	4	3	4	2	5	3	4	0	2	0	1	3	6	2	5	**9**	3

In only one example (32), one of the given translation options is entirely ruled out, i.e., no translator has chosen alternative A in example 32. Even though some examples contain zero values for A or B, these are complemented for the others with a non-zero value for EQ (in example 3, 14, 17, 21, 36, 38, 40 and 48).

The respondents were asked to state motivations for their choices and when the responses were studied, it was found that most of the motivations provided were vague and subjective. There are numerous "feelings", and value statements such "better", "clumsy", "correct" and so on. Some prefer translations closer to the

source, some want them to be more "Swedish". Some prefer more specific and some prefer more general translations. In several cases, totally contradictory motivations were given for the same example. A sample of the comments is shown in Table 23 below.

Table 23. A sample of motivations to translator choices (translated into English as the questionnarie was given in Swedish).

Sounds better, more Swedish.	*More correct.*
*Sufficient information in **A**.*	*None of the options feel Swedish*
Better style	***B** is wrong.*
Clearer, smoother.	***A** is too elaborate.*
*The clarity in **A** is sufficient.*	*The phrase **X** is to be preferred instead of the unnecessary formal **Y**.*
*In **A** the word **X** refers to two different things, yielding ambiguity.*	***A** is much clearer and easier to understand.*
***A** is clumsy, **B** is clear and "good" Swedish.*	*Maybe **B** is a little better.*
Both are clumsy.	*Better Swedish.*
***A** contains factual errors.*	*More coherent text.*
I prefer plural in headings.	*Nicer to read.*
***X** is obvious and need not be spelled out.*	*Closest to the original.*
***B** is almost too clear, even though it is not wrong.*	*I want an EXACT translation!*
It "feels" better.	*Plural is better.*
***A** is simpler.*	***X** is superfluous.*
*Should be **B** (out of old habit)*	***X** should not be translated.*
More straight-forward and more general.	***A** is more relaxed.*
"Feels" more general.	***A** is unnecessarily wordy.*
Feels incomplete because an indefinite article is missing.	*Better flow.*
*The preposition **X** is better.*	*Better word order.*

Many respondents do not give motivations at all, and the ones who do are very brief. Perhaps the motivations would have been more elaborate if this aspect of the questionnaire had been emphasized, but then some translators may not have filled in the questionnaire as it would have required considerably more time to complete. The lack of consensus when picking the "best" translation gives rise to certain questions regarding the use of translation memory software. If these translators had worked using a translation memory-based tool, and if one of the alternatives (A or B) had been presented as *the* suggestion to use, what would their reaction have been? Previously it was concluded that the general aim within technical translations is to strive for consistency, given certain constraints on context. Would they actually use the suggested alternative or would they prefer something else? Something that is "better", "more Swedish", for instance?

3.4 Summary

This chapter illustrates the way that empirical translation studies can contribute to the understanding of some fundamental concepts built into translation support tools.

The results of this study can be summarized in the following points:

1. Translators of software manuals do strive for consistency in general.

2. Translators have a positive attitude in general towards translation tools, such as translation memories, but show some hesitation as regards the change of the role of the translator when working with such tools.

3. The only explicit cause for breaking the consistency is when a repeated source segment occurs in different functional contexts, for example as a heading and as a cell in a table.

4. There are differences in attitudes towards consistency as expressed by translators, project leaders at translation companies and the translation customer. The trend is that translators and project leaders are more positive to consistency than the customer. The customer is not as negative towards variation in running text as the translators and project leaders. The reason for this is to be found in the different perspective that the three types of actors have in the translation project.

5. The choice of "best translation option" varies considerably among translators, which points to the problem of making translators accept suggested translations from translation memory-based programs.

One important fact that this study revealed is that translators are not a homogeneous group of professionals as they often have individual criteria for determining what a "good" translation should look like. What is reusable for one translator may not be reusable for another translator. When new technology is developed and introduced to a group of professionals, factors like the translators' attitudes towards tools and their perception of the the nature of translation work as a very individual skill should be taken into account. Translation memories are intended to increase efficiency and certain aspects of quality (Heyn, 1996, 1998), but if they are not introduced to professional translators in an appropriate way, the result could be that the translators will not use the tools the way they were intended.

In chapter 2, section 2.4.2, translation memory tools were discussed, and the lack of productivity gains when companies started to use tools was brought up. In addition to the explanation that the technical and administrative problems of sharing translation memories among several translators were the cause of inefficiency, we can now add the observation stated in point 5 above, namely that translators have difficulties in accepting existing translations.

4 A Parallel Corpus – The Linköping Translation Corpus

In this chapter the empirical foundation for this thesis is presented, namely the Linköping Translation Corpus, which contains several English-Swedish parallel texts and which have been created with the aid of the DAVE package (see next chapter). This chapter also presents some data extracted independently from the source and target texts: word type/token ratios and recurrence profiles.

4.1 The Linköping Translation Corpus

The Linköping Translation Corpus consists of English source texts linked to Swedish target texts. The text material comes from two major text types: user's guides to computer programs and fiction. There is also a shorter machine-translated text consisting of dialogue included in the corpus. The user's guides have kindly been given to us by Microsoft Corporation and by IBM. The Microsoft texts were translated in the traditional way, i.e., without any designated translation support tools such as translation memories, but the IBM texts were translated with the aid of IBM Translation Manager. The short text translated automatically was kindly provided by the Swedish Institute of Computer Science (SICS) in Stockholm and was translated with the SLT system (Agnäs et al. 1994). The two novels were distributed to us from Språkbanken in Göteborg and they are also translated without any computer-assistance.

From the start, the technical translations were the priority, and by having the opportunity to compare translations made using different translation methods added an extra dimension. The novels were then included in the corpus as a kind of reference material, to which the manuals could be compared. Finally the automatically translated text (ATIS) was added to the corpus to be able to compare human translation, translation memory-aided translation to fully automatic translation. In order to investigate repetitions and consistency in the translations, it was decided to include the complete texts in the corpus, instead of just samples.

Table 24 below shows an overview of the translation corpus:

Table 24. The Linköping Translation Corpus - an overview

	Text type	Source lang.	Target lang.	Title	No. of source words	No. of target words	No. of links	Translation method
1	User's Guide	English	Swedish	Microsoft Access User's Guide	179,631	157,302	14,704	Human (traditional)
2	User's Guide	English	Swedish	Microsoft Excel User's Guide	141,381	127,436	12,589	Human (traditional)
3	User's Guide	English	Swedish	IBM OS2 User's Guide and Installation Guide	127,499	99,853	11,932	Translation memory
4	User's Guide	English	Swedish	IBM InfoWindows User's Guide	69,428	53,619	7,771	Translation memory
5	User's Guide	English	Swedish	IBM Client Access for Windows User's Guide	21,321	16,752	2,426	Translation memory
6	Novel	English	Swedish	Gordimer: A Guest of Honour	197,078	210,350	12,254	Human (traditional)
7	Novel	English	Swedish	Bellow: To Jerusalem and Back	66,760	65,268	4,209	Human (traditional)
8	Dialog text	English	Swedish	ATIS dialogues	2,179	2,048	263	Automatic (MT)
Total					805,277	732,628	66,148	

The eight parallel texts included in the corpus are stored in three different representations:

1. Tab-separated text files (with the source sentence and target sentence in the same file)
2. Microsoft Access MDB files.
3. Separate source and target text files, where each record is numbered consecutively, ##1##, ##2##, ##3##…

The reason for having three parallel formats is due to the development in time. The tab-separated text files were the first format used by the alignment program, the MDB-format was necessary for efficient and convenient search in the bilingual concordance component and the last format was taken into use when work on word alignment was started. Although, the use of different formats may seem a bit awkward, in practice it does not cause any major problems as there are conversion tools to and from all formats. The main issue is that the different representations hold the same content and the same structure.

The aligned texts in the corpus also contain information on what kind of mapping relation each sentence pair has (1-1, 2-1, 1-2, etc.).

From each of the eight texts, 100 sentence pairs were randomly sampled and these pairs were tagged with information on structural and semantic correspondences. This correspondence study is described in chapter 10 and chapter 11.

4.2 Analysis of source and target texts independently

With the help of the DAVE tool package (described in chapter 5), both the source and the target texts were analyzed as separate texts. The simple parameters analyzed were the following:

- Number of running words
- Number of word types
- Word type/token ratio
- Number of sentence tokens
- Number of sentence types
- Average number of words per sentence
- Number of repeated sentences
- Recurrent sentence rate.

Without going into the parallel corpus, it is then possible to set up some criteria for comparing each source text with its corresponding target text. These criteria included:

- Number of source sentences compared to number of target sentences
- Number of source words compared to number of target words
- Recurrent sentence ratio in the source text compared to recurrent sentence ratio in target text.

Let us first start by looking the general data for the different source texts in the corpus, see Table 25 below. In the following tables, the data for the short automatic translation, ATIS, are presented together with the other texts, but as this text is not really comparable to any other texts, the data for this text will not be commented on at this stage.

Table 25. Source texts - general data

	Access	**Excel**	**OS2**	**InfoWin**	**Client**	**Gord**	**Bellow**	**ATIS**
Word tokens	179631	141381	127499	69428	21321	197078	66760	2179
Word types	4370	4483	7537	3276	1680	17539	10139	245
Word type/ token	41.11	31.54	16.92	21.19	12.69	11.24	6.58	8.89
Sentences	14829	12610	12242	7834	2427	12310	4215	263
Words/ sentence	12.11	11.21	10.41	8.86	8.78	16.01	15.84	8.29
Repeated sentences	5361	3807	3333	4116	904	184	4	0
Recurrent sentence rate	14.7%	13.62%	13.93%	31.10%	17.55%	0.18%	0.01%	0.00%

There is nothing strikingly unexpected in the above table, although it should be noted that the two novels (Gord and Bellow) contain the largest number of word types, longest sentences and lowest recurrent sentence rates. The fact that there are 184 sentences which are repeated in the Gordimer novel, may even seem somewhat high as repetitiveness is not a common characteristic for fiction, but at closer scrutiny, it turns out that almost all the repeated sentences is contained in the dialogue part of the novel. For example, utterances like "I know.", "Yes." and "All right." occur several times in the novel and constitute to a large extent these 184 repetitive sentences. Furthermore, it is worth mentioning the relative similarites between the two Microsoft texts (Access and Excel); the number of words per sentence is comparable as well as the recurrence rates. Of the IBM texts, the InfoWin text has a considerably higher recurrence rate than the others (31.1 per cent).

We can now investigate the Swedish target texts in the same way (see Table 26 below.)

Table 26. Target texts - general data

	Access	Excel	OS2	InfoWin	Client	Gord	Bellow	ATIS
Word tokens	157302	127436	99853	53619	16752	210350	65268	2048
Word types	6703	7246	10152	4308	2266	23599	13026	255
Word type/ token	23.47	17.59	9.84	12.45	7.39	8.91	5.01	8.03
Sentences	15079	13020	11943	7735	2457	13427	4285	263
Words/ sentence	10.43	9.79	8.36	6.93	6.82	15.67	15.23	7.79
Repeated sentences	5040	3853	3066	4351	933	291	8	0
Recurrent sentence rate	11.37%	13.06%	9.84	39.26%	18.70%	0.31%	0.02%	0.00%

Although the figures vary slightly, the same pattern is discernible here as for the source texts, namely that the novels contain the highest numbers of word types, longest sentences and lowest recurrent sentence rates. The Microsoft texts and IBM texts also seem to be relatively similar.

The next step is then to compare the general data from the source and target texts and see if we can conclude something about the translations. We do this by comparing the relative proportions of number of sentences, number of word tokens and recurrent sentence rates, by using the following simple measures:

ST-Sentence = the number of source sentences/number of target sentences

ST-Word ratio = number of source words/number of target words,

ST-Recurrent sentence ratio = Recurrent sentence rate(Source text)/Recurrent sentence rate(Target text).

These comparisons are summarized in Table 27 below.

Table 27. Relative comparisons between source texts and target texts

	Access	Excel	OS2	InfoWin	Client	Gord	Bellow	ATIS
ST-Sentence ratio	0.98	0.97	1.02	1.01	0.99.	0.92	0.98	1.00
ST-Word ratio	1.14	1.11	1,28	1.29	1,27	0.94	1.02	1,06
ST-Recurrent sentence ratio	1.29	1.04	1.09	0.79	0.94	0.58	0.50	N/A

The figures tells us that only two of the texts have more source sentences than target sentences (namely OS2 and InfoWin). This could indicate that most of the texts contain more deletions or that the sentence pairs have a high degree of n-1 sentence correspondences, but at this point this is mere speculation. Only one

text, the Gordimer novel, contain more running words in the translation than in the original text. Due to fact that Swedish contain more compounds (written as single words) than English, and that at a large proportion of the definite article "the" and the verb "do" do not have Swedish counterparts, it would be reasonable to expect that the number of words be smaller in the Swedish text. But, again we can only speculate that the text type, in this case fiction, seems to give rise to a relatively higher number of target words. This is apparent if we also look at the word ratio for the other novel by Bellow which contain fewer words in the translation compared to the original, but the figure (1.02) is still considerably lower than for the translations of the computer manuals. The English-Swedish Parallel Corpus (ESPC) from Lund contain comparable ST-word ratios for the translations of fiction from English to Swedish (0.98). Looking at the total material (English to Swedish) including non-fiction, gives a ST-word ratio of 1.003 in ESPC.[6] This means that the non-fiction part of the ESPC corpus contain more source words than target words (ST-word ratio 1.028). The computer manuals in the Linköping Translation Corpus do seem to be different in this respect as the ST-word ratios range from 1.11 to 1.29. In relative terms it is reasonable to expect that more information is preserved or added in the fiction translations compared to the translations of manuals.

A reasonable first hypothesis when it comes to sentence recurrence rates would be that texts translated with translation memories would have have at least as high sentence recurrence rates in the target text as in the source text. The reason that it can be even higher is that if the text was translated consistently on the sentence level, which translation memory tools will encourage the translator to do, more or less all source repetitions would be repeated in the target text as well. Furthermore, as translation memory tools also have the capacity to find fuzzy matches, the translator would be steered towards an even higher degree of consistency in the target text. Another hypothesis is that translators who do not use translation memories or any other computerized tools, will find it harder to identify repetitions in the source text, which leads to a less consistent target text. It should therefore be logical to expect that human (traditional) translation of technical texts would result in lower sentence recurrence rates in the target texts than in the source texts.

The figures for ST-Recurrent sentence ratio in Table 27 should be interpreted as follows: if the value is exactly 1, then the source and target text contain the same proportion of recurring sentences. If the value is higher than 1, then the source text has higher sentence recurrence rate than the the target text. Consequently, a value below 1, shows the opposite, namely that the target text is more recurrent. When we compare the values for sentence recurrence in the texts, we can see that two of the IBM texts (InfoWin and Client) actually have higher sentence recurrence rates in the target than in the source, which is in line with the first hypothesis as these texts were translated with the aid of translation memories. The two Microsoft texts have higher recurrence rates for the source text than the target text which is in accordance with the second hypothesis,

[6] Data from ESPC was kindly provided to me by Bengt Altenberg (personal communication October, 7 1999). The fiction ST-word ratio is based on 276,591 source words/281,127 target words and the total ST-word-ratio on 494,374 source words/492,885 target words.

namely that consistency on the sentence level would be more difficult in traditional translation. The text that does not fit the pattern then is the OS2 text, which has a higher recurrence rate in the source text than in the target text even though the translation was produced with translation memory. The two novels (Gordimer and Bellow) do not contain enough recurrent sentences to be interesting in this respect.

4.3 Recurrence in source and target texts

If a source and a target text are analyzed independently for repetitions, the result may yield that twenty per cent of both the source and target text are repetitive on the sentence level. However, does the similarity of recurrence degrees mean that the source and target repetitions actually correspond? Let us look at a source and target texts to see how well the most frequent sentences correspond in numbers. Five of the most frequent sentences in the English computer manual for OS/2 are shown in Table 28.[7]:

Table 28. Five frequent source sentences of OS/2 User's Guide

Sentence	FRQ
Click mouse button 2.	80
Open OS/2 system.	69
Open system setup.	42
Press and hold mouse button 2.	39
Select the arrow to the right of Open.	34

If we expect a strong isomorphic relationship between repeated sentences in the source and target text, there should be similar frequency ratios in both texts for corresponding sentences. However, none of the top ten sentence types have been consistently translated in the Swedish translation. By looking at the sentence frequency lists for both the source and the target, it is not difficult to identify the correspondences, but the differences in frequency are striking. An example of such differences is shown in Table 29.

Table 29. Correspondences between repeated sentences in the source and target

Source	FRQ	Target	FRQ
Click mouse button 2.	80	Klicka med musknapp 2.	18
Open OS/2 system.	69	Öppna systemprogram.	43
Open system setup.	42	Öppna systemkonfiguration.	38
Press and hold mouse button 2.	39	Håll ned musknapp 2.	16
Select the arrow to the right of Open.	34	Välj pilen till höger om Öppna.	30

[7] The examples are taken from the ten most frequent text sentences, excluding headings and table cell texts which often only consist of one or two words.

There are of course translations that are less inconsistent than the example above. In Table 30 below examples of frequency correspondences from a Swedish translation of Microsoft's Access User's Guide are given.

Table 30. Examples of repeated source sentences with corresponding translations in Access UG

Source	FRQ	Target	FRQ
Choose OK.	37	Välj OK.	39
From the Toolbar shortcut menu, choose Customize.	9	Välj Anpassa på snabbmenyn för verktygsfält.	9
Open the report in design view.	7	Öppna rapporten för design.	6
Open a database in Microsoft Access or switch to the database window for the open database.	6	Öppna en databas i Microsoft Access eller växla till databasfönstret för aktuell databas.	4
Choose the New button.	6	Välj Ny.	5

The differences in frequency between source and target are considerably smaller here than in the OS/2 translation. However, a superficial comparison of the lists of recurrent sentences and their frequencies, would not be sufficient to tell how consistently the translations had been made. The Discrepancy tool which will be described in detail in the next chapter (section 5.4), can assist us in finding out exactly how the recurrent sentences are related to each other, but as this involves analysis of the parallel texts as one object, this analysis will have to wait until chapter 6 (section 6.2).

4.4 Summary

By using simple tools and only analysing very general data from the source and the target texts independently, a translation scholar or a translator can still make a number of observations from the texts. For example, a translator could immediately decide, by only looking at the source text, that the use of translation memories would indeed be a waste of effort for the novels in this corpus. However, given the recurrence rate for the technical texts from Microsoft and IBM, using translation memories would be a tempting option.

As the texts are not aligned at this stage, a translation scholar would not be able to look at the source and target text in parallel, but analysing the source and target texts independently and comparing the data for each source text with its target counterpart will reveal some basic data about the relationship between the originals and the translations.

To be able to make more detailed observations on the relationships between the source and the target text, it is necessary to investigate the source text and target text as a whole, that is, as parallel texts, and this is the focus of chapter 6. However, first the DAVE toolbox will be introduced in the next chapter.

5 A set of translation corpus tools

In chapter 2, different types of tools related to translation and translation corpora were discussed. In this chapter a set of such tools will be presented. The tools have all been implemented, in the first stage as independent programs with rudimentary interfaces (operated from the command line in DOS and Unix) and in the second stage the tools have been integrated into one Windows system with a graphical and more user-friendly environment, nicknamed DAVE[8] (Diagnosis, Alignment and Verification for the Editor).

Before describing the tools, certain terms and use of terminology have to be explained:

- A *sentence* is a string of words separated by punctuation marks and upper-case letters or separated by carriage return and tab characters. The use of the term *sentence* will include both text sentences and technical sentences (headings, table cells, list items, etc.)
- A *multi-word unit (MWU)* is a sequence of two or more consecutive words that does not cross sentence boundaries.
- A *recurrent multi-word unit* is a segment that occurs in at least two different sentences of a given text or corpus. For practical reasons, a higher lower bound than two is often used for the number of occurrences, but there is no non-arbitrary way to fix this lower bound. Note that a recurrent segment as such may be a proper syntactic unit, or a collocation, but it can be neither.
- A *phrase* is a multi-word unit constituting a syntactic unit with compositional meaning.
- A *collocation* is a multi-word unit constituting a lexical unit of some sort.

The DAVE package contains the following components:

[8] The DAVE package was converted to a graphical environment from the command line tools (Merkel 1992, Merkel et al. 1994, Merkel 1996) by Mikael Holm and Mathias Olsson. The conversion task is described in their Master's thesis (Holm and Olsson 1996).

1. Extraction of recurrent units
2. Diagnosis of recurrence
3. Paragraph Alignment
4. Sentence Alignment
5. Discrepancy Analysis
6. Bilingual Concordancing

The Extraction component contains functionality for retrieving recurrent units on the sentence, multi-word unit and word levels. The Diagnosis component provides tools for calculating how recurrent a given text or set of texts are on different levels. With the Paragraph and Sentence Alignment components, the user can build a parallel text where the text is synchronized on the paragraph and the sentence levels. The Discrepancy analysis provides information on how consistent or inconsistent the translations of the recurrent sentences are. Finally, the Bilingual Concordance program lets the user browse and search the parallel text for any combination of source and target words and MWUs and collect data for the co-occurrence of certain items.

5.1 Extraction of recurrent units

When a translator or a translation team is confronted with a source text that is going to be translated, there is usually no information available about the characteristics for the text. To make the right decisions regarding translation method, selection of translators with certain skills and what existing resources can be used during the translation process, the translation team needs to discover characteristics about how repetitive the text is, if it is similar to other previously translated source texts, and the range of vocabulary and terminology that needs to be standardized before the translation work starts. The writers of the source text may have reused portions of the text in different parts of the source text, they may have standardized certain constructions and stylistic conventions, but it is more common that such information never reaches the translators. So to make the right decisions the translators need to acquire this type of information themselves.

DAVE contains diagnostic features that can help the translators to find out more about the nature of the source text. The diagnostic nature of the tool consists of the following:

- Data on how repetitive the source text is in terms of repeated technical sentences (sentence recurrence rate)
- Data on how repetitive the source text is in terms of repeated multi-word units within technical sentences
- Data on how similar the source text at hand is to a previously translated source text (comparing sentences and repeated multi-word units)

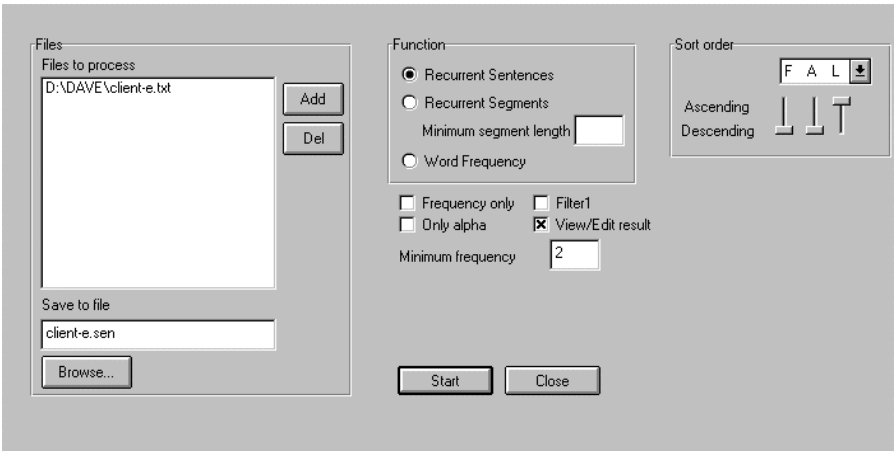


Figure 9. Dialog box from DAVE for extracting recurrent sentences, multi-word units or words

Before the actual diagnosis takes place, the relevant units of the text have to be extracted, for example sentence types, multi-word units and single words. This is done in the Extraction part of the tool. The dialog box in Figure 9 shows how the translator can adjust the options for the extraction process.

The options selected in Figure 9 will produce a list of all sentences in the input file that occur twice or more. If the translator would like to extract multi-word units, the button "Recurrent Segments" would have to be selected and the minimum MWU length (in number of words) specified. Sentences are identified in a technical sense, i.e. they are identified by punctuation information, tab and carriage return characters.

Retrieval of recurrent sentences

Input: Text
Output: A segment table with sentence types (as word strings), number of occurrences for each sentence in the text and position (document identifier and a list of offsets).

The output from a typical retrieval of recurrent sentences is illustrated in Table 31 below:

Table 31. Example of recurrent sentences with frequency and file positions

Sentence	FRQ	File and position
/.../		
double-click on the configuration icon in the client access/400 for windows 3.1 group window.	6	client.eng 10801 12687 16272 17748 20041 61801
hold down the left mouse button and move the mouse pointer to the left to make the column narrower and to the right to make it wider.	5	client.eng 93955 102615 107684 111562 113677
click on the item with the left mouse button, or move the cursor up or down to highlight the item.	5	client.eng 91736 92420 96111 103797 104290
position the mouse pointer on the column separator line to the right of the column heading you want to change.	5	client.eng 93773 102433 107502 111380 113495
select the save or save as option from the file pull-down menu to save the session profile.	5	client.eng 80568 81532 82355 83878 84735
a double arrow is displayed when the mouse point is on the line.	5	client.eng 93885 102545 107614 111492 113607
to change the width of a column:	5	client.eng 93735 102395 107464 111342 113457
close the client access configuration window.	5	client.eng 13112 13642 14900 16991 17553
/.../		

5.1.1 Retrieval of recurrent multi-word units (Frasse-1)[9]

Recurrent multi-word units are retrieved from the text after it has been split up into sentence types. Either you can input the text (a table with sentence types will be created, see Table 31) or one or more existing segment tables. The data structure for storing multi-word units consists of a table of multi-word units. Each entry in the table contains the following fields:

- *The multi-word unit*
- *A list of locations* where each location contains a document identifier and a list of offsets inside each document.
- *A parsed word string*, a sequence of references to the word table, see below. All text segments are first parsed to word strings so that comparisons of words can be done by just comparing pointers. (This field is not present in the output.)

In addition to the above fields, the data structure of the segment table holds a *word table*, i.e. a table of known words with frequencies.

[9] Frasse-1 has previously been described in Merkel, Nilsson and Ahrenberg (1994).

Each recurrent multi-word unit is stored in only one entry, but has multiple locations with multiple offsets.

Input: Text or a list of multi-word unit tables

Output: A new multi-word unit table holding maximal multi-word units from input.

All combinations of multi-word units from the input table are compared pairwise. Each pair, <*s1*,*s2*>, is searched for common parts. A common multi-word unit is defined as all multi-word units that contain the same consecutive words in both *s1* and *s2*. If the identified segment is long enough, it is stored with all locations and adjusted offsets from both *s1* and *s2*.

Only multi-word units that start at locations are considered, such that the immediately preceding words, if any, are different in *s1* and *s2*. A multi-word unit ends in the same manner. This means that only the longest possible multi-word units are considered. Note that both *a b c d* and *a b c d e* can be regarded as maximal segments if they have different frequencies, i.e. if *a b c d* occurs 15 times and *a b c d e* 12 times they are both maximal multi-word units.

The fact that only the longest possible multi-word units for each pair are stored has the effect that there is a significant reduction of data, which makes the algorithm more efficient.

There is a filter that trims the multi-word units at the head and tail of the multi-word unit, which reduces the size of the multi-word unit table. The filtering strategies are discussed later in this section.

The resulting multi-word unit tables can be sorted in various ways (any combination of sorting by alphabetical order, length of units and frequency). It is also possible to strip away the location information, which leaves a list of segments and their corresponding frequency data. The stripped multi-word unit list is easier to handle when a translator is viewing or revising the units. After revision it is possible to extract the location information for each multi-word unit in the revised multi-word unit list and create a new complete MWU table.

Table 32. The top 32 MWUs generated from a computer program User's Guide (without filtering)

Multi-Word Unit	Freq	Multi-Word Unit	Freq
you want to	452	and then choose	82
, you can	392	the database window	80
for example,	327	in this chapter.	79
menu, choose	136	, click the	79
if you want	130	the edit menu	78
you can use	119	you can also	78
to create a	112	the tool bar	77
, and then	109	a form or	73
example, you	106	in design view	73
if you want to	105	choose the ok button	72
, see chapter	105	you can create	72
for example, you	102	the ok button.	71
in this chapter	100	, select the	71
the qbe grid	99	then choose the	70
form or report	94	choose the ok button.	69
, see "	88	for more information	69

In Table 32 the 32 most frequent maximal MWUs are listed. The text which was analysed was a 100,000 word user's guide for a database program. MWUs consisting of 3 words or more were searched for without the use of any filters. The result was a list of maximal strings in the source text of which many are useless as translation units.

When this kind of output from the system was revised by hand, it was found that a majority of the MWUs removed from the original output actually were units that contained punctuation marks, ended with phrase-initial function words (such as "and", "the", "to", "in", etc.) or had only one parenthesis character or quotation mark. Therefore a filtering module was implemented that made it possible to define words that should be stripped at the beginning and at the end of MWUs as well as requirements on what kinds of characters should be regarded as pairs (quotation marks, parentheses, etc). Table 33 below shows the result of running the system on the same text as in table 4 with the filtering mechanism on. Of the 32 most frequent segments in the unfiltered result above, 22 have been filtered out, leaving a residue of 10.

Table 33. The top 10 segments generated by the extraction module (with filtering)

Multi-Word Unit	Freq
in this chapter	100
the qbe grid	99
form or report	94
the database window	80
the edit menu	78
the tool bar	77
the ok button	74
in design view	73
choose the ok button	72
for more information	69

The filter can be tailored to specific texts, text types and languages by simply editing a word list. In the above example, the most frequent segments are either prepositional or noun phrases due to the characteristics of the used filtering specification of function words. A simple filter like this will, of course, not filter out all non-interesting MWUs, but it will reduce their numbers significantly.

An extract of a filter file is shown below:

```
Pair: “ “ ( ) { } [ ]
skip_before: , . ; :
skip_after: , . ; : ?
skip_after: the a an for in on is are each every want can related might click you’re
skip_after: or and one also to too its save associate associated relate deleting
skip_after: if this that at you from of then your not do does did doing
skip_after: when how create creates many several design designs designing most
skip_before: the a if when how your his her my then and or not do did
/…/
```

The filter specifies that no MWU can begin or end with a punctuation character, nor can it end in any of the words specified by the label “skip_after:”. Furthermore an MWU can not begin with any of the words given after the label “skip_before:”.

The extraction module described in this section can be made more powerful in several ways. A statistical measure, which is often used as an indicator of collocations, is mutual information (Church and Hanks 1990). Smadja (1993) used the Dice coefficient for finding both monolingual and bilingual MWUs, and Shimohata et al. (1997) explored the possibilities of retrieving MWUs by calculating entropies for different left and right word contexts, where co-occurrence and constraints on word order will help to single out likely candidates for useful MWUs. There are studies, however, that have indicated that high frequency is a stronger indicator for retrieving MWUs than mutual information (e.g. Daille 1994), which would support the frequency approach taken in the described MWU extraction component of Dave.

Another path to improve the quality of MWU extraction is to use knowledge of linguistic structure (Kupiec 1993). The identification of phrases can be made more precise by shallow parsing of the output on the basis of punctuation marks and function words such as articles, prepositions and conjunctions. Such forms mark boundaries of minimal segments (McDonald 1992) and it can be postulated that the majority of useful phrases have their boundaries coinciding with the boundaries of these minimal segments (a recurrent phrase may span more than one minimal MWU, however).

5.1.2 Combining filtering and entropy thresholds to retrieve multi-word units (Frasse-2)

To test and characterize the differences between a statistical approach, such as the one described in Shimohata et al. (1997), with the frequency approach described in the previous section, a new MWU extraction component was built where the word filtering approach and constraints on entropy thresholds for the immediate contexts of MWU candidates were combined. Furthermore, it was assumed that by changing the algorithm slightly processing could be performed more efficiently.

Shimohata et al. (1997) describe an algorithm which is based on the observation that most MWUs appear in highly varying contexts. The words just before and after an MWU vary a great deal, while the actual MWU stays the same. The diversity of neighbouring words thus marks where MWUs start and end. The entropy value for an MWU is then a combination of the entropy values measured to the left of the MWU and to the right of it.

To measure the probability of each adjacent word ($w_1 \ldots w_n$) to a given string of words (*str*) the relative frequencies are used in the following way:

$$p(w_i \mid str) = \frac{freq(w_i)}{freq(str)}$$

The left and right entropies of a string of words, *str*, are mathematically defined as:

$$H(str) = \sum_{i=1}^{n} - p(w_i) \log p(w_i)$$

where $p(w_i)$ is the probability of seeing the word w_i adjacent to the string, and w_i are the words that do occur just before or after the string in the text.

A high entropy value signifies that the words surrounding the string, *str*, varies considerably, so strings with

$$H(str) \geq T_{entropy}$$

are accepted as phrases. The entropy threshold used by Shimohata et al. (1997) was set to 0.75. This threshold is then used in conjunction with a frequency threshold which indicates the minimum frequency of the string.

The approach to use entropy values as a constraint for extracting MWUs was then combined with a modified version of the filters described in the previous section. The reason for this is that when purely statistical approaches are used, a lot of meaningless recurring patterns such as "for a", "in the", "out of the", etc. are extracted as likely candidates for MWUs. In the previous filter, there were in principle two ways of constraining the extraction of such MWUs: (1) a list of

words that could not start an MWU; and, (2) a list of words that could not end an MWU. In evaluations of the output from the first MWU extraction component, it was observed that words listed as non-starters and non-enders could very well be included inside an extracted MWU, but that there was no way to constrain the extraction component to do this. For example, if the user decides that personal pronouns such as "you", "he", "him", etc. should not be part of meaningful MWUs in a technical domain, then there should be a way of expressing this. Consequently, a third category of filtering entities was added: *prohibited words*, i.e., words that are forbidden to be included in MWUs. A fourth category of filtering entities was also added: *ignored words*. The latter are words that are ignored (or skipped) when entropies for surrounding contexts are measured. In some of the English applications the definite article "the" was included in this category.

The proposed language filter for the second version of the MWU extraction component (from here on called "Frasse-2")[10] consists of

1. a list of words that MWUs may not start with (NON-STARTERS)
2. a list of words that MWUs may not end with (NON-ENDERS)
3. a list of words, which can never be part of any MWU (PROHIBITED WORDS)
4. a list of words that are not considered part of an MWU, but do not delimit them

In addition to these word lists, a list describing punctuation characters, which are never part of MWUs in the language in question is used.

5.1.2.1 Algorithm

The algorithm works in the following way:

1. Read the text into a two-dimensional array and store all words in a hash table together with their positions in the text.
2. Find all allowed MWUs up to a limit of L words by expanding candidate MWUs to the right by 1 word (in the first iteration all words taken as MWUs of length 1 are tested, with the exception of words below the frequency threshold, words listed as non-starters and prohibited words).
3. For each of the allowed MWUs found in step 2, calculate the left and right context entropies and store all MWUs with an entropy value higher than the specified entropy threshold.
4. Go back to step 2 and expand the stored MWUs by 1 word. This is done by finding the position of the N-gram from the hash table, and extracting the

[10] The first MWU extraction version is also described in Merkel et al. (1994) and as that first piece of extraction software was called *Frasse*, lack of imagination resulted in calling the second version *Frasse-2*.

next word position from the text array. The iteration then continues until no more MWUs can be added or until the limit of the longest MWU (specified as L) is reached.

In the second phase all the found high-entropy MWUs are examined and text positions where shorter MWUs are subsumed by longer ones are deleted. This results in a list of MWUs that meet the frequency threshold, the entropy threshold and are constrained by the language filter. The output is then listed in two formats: (1) an alphabetically sorted list of MWUs with information of the entropy value and frequency as well as the corresponding text positions; and (2) a list of MWUs sorted by size and in descending entropy order.

The differences between the output from Frasse-2 when the word filter is used and when it is not, are illustrated in Table 34 (the first ten entries starting with the letter L).

Table 34. Ouput from Frasse-2 with and without filtering

With filter	Without filter
label controls	label after you
label tool	label and
last field	label controls
last name	label for
last name field	label for the
last names	label from
last page	label from a
last record	label in the
layout properties	label of
layoutforprint property	label of the

Frasse-2 has not been included in the DAVE package yet. Instead it has been implemented in Perl as a stand-alone program which has mainly been used in conjunction with the word alignment system (LWA) described in chapter 7.

In the next section, the two different versions of the MWU extraction programs (Frasse-1 and Frasse-2) will be compared and evaluated.

5.1.3 Comparison of the first and second version of the MWU extraction programs

What are the results in terms of output data and efficiency if the two different approaches to extraction of MWUs described in sections 5.1.1 and 5.1.2 are compared? In the first approach, only frequency data and a language filter were used to constrain the different n-gram units (Frasse-1). In the second approach (Frasse-2), frequency thresholds and a modified version of the language filter were combined with a statistical entropy measure that calculated the probabilities for adjacent strings to be considered as well-formed MWUs.

In the following evaluation, the same text has been used for both systems, namely the English Microsoft Access User's Guide, described in section 4.1. The text contains 179,631 words. The same language filter was used in both systems,

with the exception of 10 prohibited words which were only used in the second version of the system. The systems were run in two configurations, one with the frequency threshold set to 4 occurrences and one configuration with a frequency threshold of 2. For Frasse-2, the entropy threshold was set to 0.75. The minimum size of the MWU was 2 words for both the systems, and the maximum size, L, was in Frasse-2 set to 10 (in Frasse-1, this upper limit cannot be specified).

Recall cannot be measured as there is no reliable way to determine how many well-formed (or practically usable) MWUs there are in a given text, but comparisons can be made between the systems and thus relative recall can be measured. Precision has been tested on all the output of all the MWUs starting with the letter L. Here the MWUs have been judged as either "good" or as "bad" in a broad sense. A "good MWU" is defined as a well-formed lexical unit (term, phrase or collocation) and those which do not belong to this class are simply judged as "bad".

Table 35. Comparison of Frasse-1 and Frasse-2

System	Frequency threshold	Total extracted MWUs	MWUs in sample (letter L)	Good	Bad	Precision
Frasse-1	4	2655	55	43	12	82.08
Frasse-2	4	953	29	29	0	100.00
Frasse-1	2	7516	193	114	79	59.06
Frasse-2	2	1880	49	49	0	100.00
Frasse-2 (no prohibited words)	2	2,850	65	58	7	89.23

The strengths and weaknesses of the two versions are obvious from the above table. Frasse-1 will produce a higher number of MWUs, but of lower quality. Frasse-2 on the other hand extracts fewer MWUs (low recall), but does this with 100 per cent accuracy. In order to increase the recall of Frasse-2, all prohibited words were deleted from the language filter, which makes Frasse-2 behave more like Frasse-1 as this feature is not included in the first system. Without prohibited words in the language filter, Frasse-2 extracted 2,850 MWUs in total, and 65 that started with the letter L, given 2 as the frequency threshold. However, precision then declined to 89.23 per cent, as seven of the 65 examples were judged as ill-formed MWUs.

To illustrate the difference between the two systems, a listing of the first 20 MWUs starting with the letter L extracted by Frasse-1 and Frasse-2 are shown in Table 36 below. The MWUs considered to be ill-formed are italicised.

Table 36. Examples of MWUs extracted by Frasse-1 and Frasse-2

Frasse-1		Frasse-2	
MWU	**Freq.**	**MWU**	**H(str)**
label control	3	label controls	3.04
label controls	9	label tool	3.04
label from a control, and later	*3*	last field	2.93
label of any size, by clicking	*2*	last name	7.24
label of the text box	2	last name and first name fields	3.16
label that microsoft access sizes as you type, by clicking	*2*	last name field	4.64
label to start and then typing the text for the label	*2*	last names	1.83
label to start, dragging the pointer until the label is the size you want, and then typing the text in the label	*2*	last page	1.83
label to the control	*2*	last record	4.9
label tool in the toolbox	3	layout and formatting elements	3.16
label tool to create a label, the label is freestanding -	*2*	layoutforprint property	3
label tool	8	levels of subforms	2.32
label, and resize the section	*2*	limit records	3.71
label, double-click	*6*	limittolist property	4.47
label, it's no longer attached to the text box	*2*	line item	2.89
label, text box	*2*	line of text	2.24
labels in the page header	3	lines and rectangles	3.84
labels into the page header	*2*	*link data*	*2.71*
last dialog box	8	linkchildfields property	1.94
last dialog box, click the finish button to display	*3*	linked object	5.14

Because prohibited words are used by Frasse-2 when it exracted the MWUs shown above, many of the unwanted MWUs extracted by Frasse-1 are filtered out. For example, the words "a", "the", "that", "you", "it's" and "to" are included in the prohibited words. On the other hand, because of the lower recall of Frasse-2, some of the well-formed MWUs found by Frasse-1 cannot be identified by Frasse-2, primarily because these MWUs do not get over the entropy threshold. It is worth pointing out that if prohibited words are used by Frasse-2, the MWUs extracted will naturally be shorter, but it does not necessarily mean that the longer MWUs that Frasse-2 produces are missed altogether. Frasse-2 will "break" an MWU as soon as a prohibited word appears, but the strings before and after that prohibited word may very well turn out to be acceptable MWUs. Consider an MWU such as "linking fields in the main report" extacted by Frasse-1. Although Frasse-2 could not detect this particular MWU, two of the subsumed strings are passed as acceptable MWUs by Frasse-2: "linking fields", and "main report". In some applications, like word and phrase linking, it is more likely that successful links can be made between the shorter subsumed MWUs and their corresponding target units, as these will have higher frequencies than the longer MWU. However, each system has its unique strength. For automatic processing, when there are few or no resources available for manual revision of

MWU data, Frasse-2 would ensure that the MWUs being extracted are of high quality. Furthermore, Frasse-2 is better at detecting more terminology-style MWUs as too long MWUs are avoided unless they pass the language filter and entropy thresholds. To illustrate that Frasse-2 does not avoid longer MWUs altogether consider the following six-word MWUs extracted by Frasse-2:

> northwind employee sales by country report
>
> employee sales by country 2 report
>
> sales by sale amount 2 report
>
> summary of sales by year report
>
> last name and first name fields
>
> northwind alphabetical list of products report
>
> order id and order date fields
>
> international section of windows control panel
>
> report or report section property setting

For other purposes, when a human can revise the more imprecise data from Frasse-1 and delete unwanted MWUs, Frasse-1 might be the system to use. Another option is to use both systems; first Frasse-2 to acquire a high-quality core of MWUs and then add the output from Frasse-1 after human revision

In practical tests, the systems are both efficient as far as time goes. Using a 500 Mhz Pentium III PC with 256 MB of RAM, Frasse-1 needed 7 minutes to extract MWUs on a 179,631 word text (min. frequency 2). The same configuration (and a 1.0 entropy threshold) took Frasse-2 just under 5 minutes on the same machine.

The main advantage for these two systems is that they do not require any linguistic information, apart from what the user would like to insert in the language filters. This means that adaptations to new (western-European) languages are easy to do (the system has been tested for English, Swedish, German and French). The use of shallow data and statistics is therefore its strength but also its weakness, as it is difficult to generalise the language filters by using word categories or grammar rules. With parts-of-speech tagged texts, it would be possible to specify the language filters in grammatical terms, and apply them together with the statistical machinery. Furthermore, Shimohata et al. (1997) and Smadja (1993) have shown how significant n-grams, i.e. uninterrupted MWUs like the ones the Frasse systems extract, can be expanded to interrupted phrasal templates, such as "For more information on …, refer to the … manual …". This is one interesting feature that could be included in this system.

Frasse-2 has been used together with the Linköping Word Aligner (LWA) in order to align texts on the word and MWU levels and to compile bilingual lexicons (see chapter 7).

5.2 Measuring recurrence

The Analysis component of DAVE can measure the recurrence rates for a text, both at the sentence level and at the MWU level. This is the component that was used in section 4.3 in the previous chapter. Given a Recurrent Sentence Table the sentence recurrence rate for the text can be measured:

> **Input**: A Recurrent Sentence Table with recorded positions
>
> **Output**: (1) Total number of words in the analysed text; (2) number of words that were part of recurrent sentences, and (3) the ratio of (2) to (1) expressed as a percentage.

Given an MWU table the MWU recurrence rate for the text can be measured in a similar way.

> **Input**: A MWU Table with recorded positions
>
> **Output**: (1) Total number of words in the analysed text; (2) number of words that were part of recurrent MWUs, and (3) The ratio of (2) to (1) expressed as a percentage.

The tool calculates the recurrence rate by positioning all MWU types on top of the actual text. The fact that MWUs may be overlapping is represented in the resulting figures. The Analysis component of DAVE contains functions for calculating the recurrence rate for one or several MWU tables. It is also possible to create a larger MWU table by unifying two or more MWU tables and record string and position data in a new MWU table. This can be useful if separate analyses have been made of texts that are related in some way. In the same way, it is possible to extract the common MWUs or sentences from two texts by using the Intersection option. This results in a new table that contains only the set of common sentences or MWUs from two ore more tables holding recurrent data.

By allowing the user to manually filter an MWU list, the quality of the MWU data can be increased. In order to be able to measure the recurrence rates, the user can rerun a manually revised list of MWUs against the original MWU list by using the option "Selection" (shown Figure 10). This process will retain the frequency and position data for all the MWUs present in the revised list, and omit any other item.

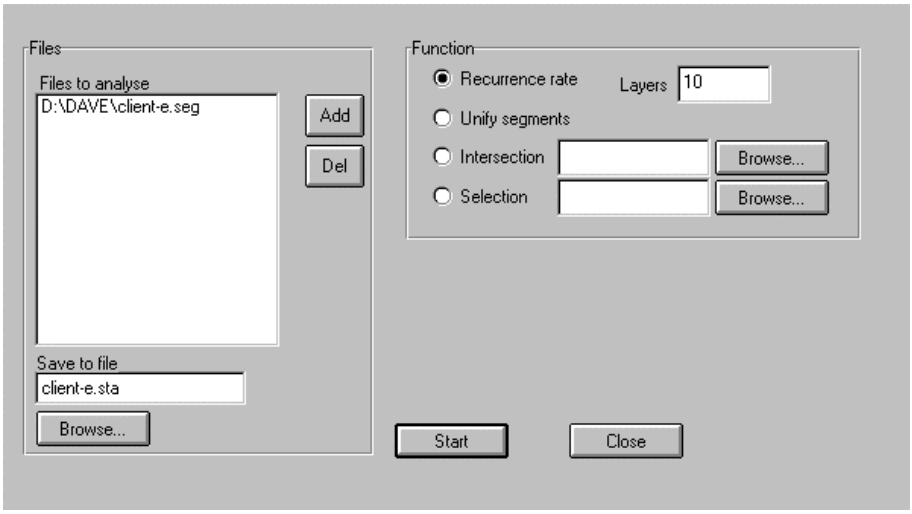


Figure 10. Dialog box with options for calculating recurrence rate, unifying MWU tables, analyzing the intersection set of two MWU tables.

An example of the recurrence data returned by the tool is shown in Table 37. The text it has been applied to here is the Client Access for Windows from IBM, which is a comparatively short text, but the recurrence rates are still relatively high.

Table 37. Example of recurrence rate

	Sentences	**MWUs**[11]
Total number of words	21321	21321
Covered number of words	3741	6869
Covered part	17.55%	32.22%

The Analysis component can be used diagnostically in several ways. For example, a text profile can be generated showing what parts are covered to what extent by recurrent items at various levels of abstraction, and the recurrent items can be checked for counterparts in an existing translation memory. Both kinds of information are relevant for deciding what efforts and resources are needed for the translation of the given text body. It can also provide information about how parts of a document are related. In a translation situation, it may be the case that a handbook is composed of chapters with different translation profiles, which could indicate that maximal translation efficiency would be gained if similar chapters were processed by one type of method, and other chapters by another method.

The MWU and sentence lists generated by DAVE can be viewed manually or analysed automatically in order to detect unnecessary stylistic variation, e.g. the

[11] In the example, the minimum frequency has been set to 4 and the minimum segment length to 3.

prepositions that go with the complements of certain nouns and verbs in examples such as *information on* vs. *information about*. A complete description of terminological variation can probably only be obtained on the basis of semantic tagging or an existing terminological database with a complete coverage of synonyms. The attempt here is to use the corpus in order to identify units that have the same translation, concentrating on concepts and constructions where consistency is vital, e.g. in the use of domain terms.

5.3 Sentence and paragraph alignment

In order to reuse translated material, there must be tools and methods to build up translation memories from existing translations (see section 2.4.2 in Chapter 2). This is called *aligning* the source and the target texts.

The alignment component in DAVE is based on the alignment algorithm presented in Gale and Church (1991). The system creates bitexts from a source and a target text, that is, it links a sentence in the original with a corresponding sentence in the target document. Apart from 1-1 relations, the program also handles 1-2 and 2-1 relations (1 source sentence - 2 target sentences, 2 source sentences - 1 target sentence). The first version (command line version) was a fully automatic alignment (i.e., the system performed all the alignments in one batch, with no possibility for the user to interact with the program, cf. Merkel 1993). This character-based system performed reasonably well on smaller sized texts. A test run on a sample of manually translated text, showed that out of 624 sentences, it failed on only 4 sentences when run in fully automatic mode. However, to correct the mistakes made on larger texts was a time-consuming and difficult task, if the aim was to produce high-quality alignments that could be used for later linguistic analyses. Therefore, the next version, included in DAVE, contained a graphical, interactive method of alignment where the alignment is performed in two steps: (1) paragraph and (2) sentence alignment.

5.3.1 Paragraph alignment

The input to the paragraph alignment component is a source text file and a target text file. The objective is to produce two new files where each file is synchronised on the paragraph level. To be able to do this the user is presented with the interface shown in Figure 11:

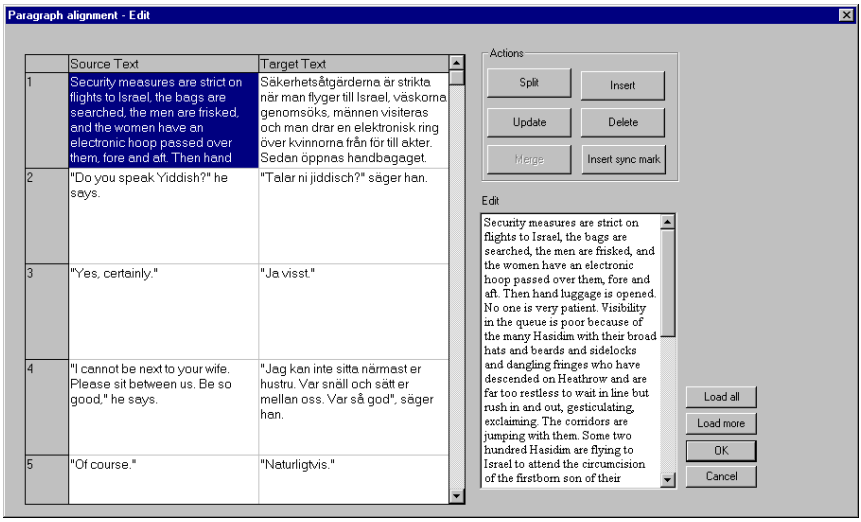


Figure 11. Paragraph alignment component

The source and target texts are presented in a scrollable table window to the left, which can be edited with the use of the Action buttons to the right. The possible actions that can be applied are (1) to split a paragraph (to make two paragraphs from one original paragraph), (2) to insert a new paragraph (empty), and (3) to delete a paragraph. Actions 2 and 3 should be seen as substeps in creating the correct paragraph alignment, for example, if the user has split one paragraph it may be necessary to delete an empty paragraph to achieve the correct paragraph alignment. The Update button is used in conjunction with the Edit window (which holds the active paragraph) and allows the user to correct any possible mistakes in the input (for example, OCR mistakes). When such mistakes have been corrected in the Edit window, the user clicks the Update button and the table to the left will be updated. To make it easier for the user, the system loads 100 paragraphs at time, and after verifying these paragraphs, the user can ask the system to load the next 100 paragraphs, or alternatively the rest of text could be loaded into the table.

If the texts are of good quality, the paragraph alignment phase is usually very straightforward. In the Linköping Translation Corpus, which was aligned with this tool, there are only minor differences between the paragraphs in the source and target texts. The maximum time for aligning paragraphs for the texts included in the Linköping Translation Corpus described in Chapter 4 was half an hour.

5.3.2 Sentence alignment

The second step towards a complete alignment is the sentence alignment, based on the paragraph-aligned input files. To guarantee high-quality results the user can choose between three different modes of alignment: (1) *Fully automatic*, (2) *Semi-automatic* and (3) *Manual* alignment ("Less Automatic"). After completion of the alignment, no matter what mode is used, the user always has the possibility to revise the alignments made by the system in a designated editor. *Fully automatic* alignment means that the system completes the alignment as well as it can. *Semi-automatic* alignment means that the system halts at error-prone passages, for example, when there are several possible ways of mapping the sentences in a given paragraph. The user can then decide if the system's proposal should be accepted or change the alignment mapping according to her own judgements. In the *Manual* alignment mode, the user has to verify the suggested sentence mappings from the system as the system halts at each paragraph. In Figure 12 the interface for the sentence alignment component is shown. Here the alignment has been run in semi-automatic (interactive) mode, and the system has highlighted a passage in paragraph 23 of the text, where the user is asked to verify the alignment of sentence number 6 where the system has suggested a 1-2 mapping.

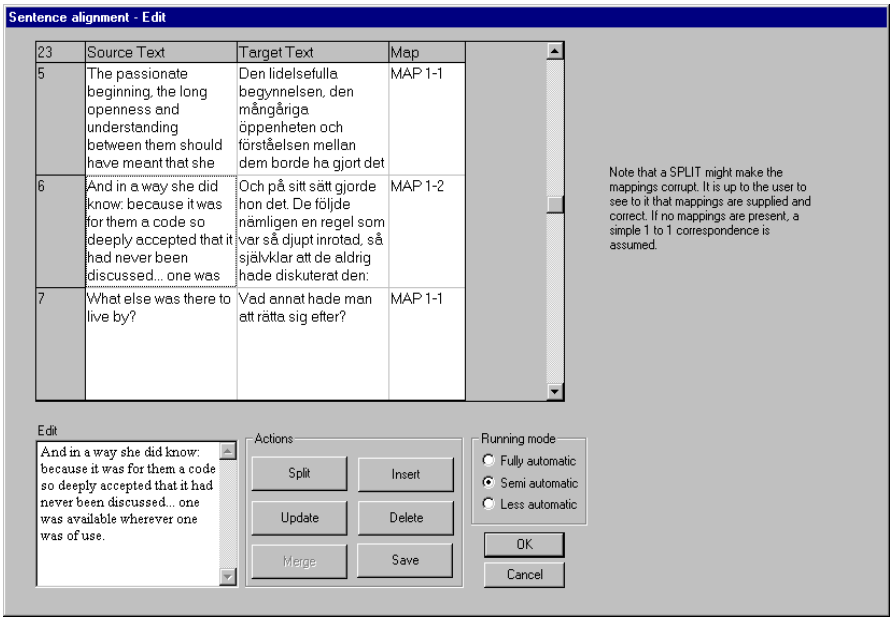


Figure 12. Sentence alignment component

The action buttons are similar to the ones used in paragraph alignment, namely Split, Insert, Update and Delete.

The experience with the alignment tool has been positive. Given a correct paragraph alignment, the time for aligning a 200,000-word corpus is about one to four hours depending on how many 1-0, 0-1, n-1 and 1-n mappings that are needed to produce the correct parallel text. Usually the system performs well on all 1-1, 1-2 and 2-1 mappings, but experience has shown that it is safer to always run the system in "semi-automatic" mode to ensure more or less 100 per cent correct output data.

5.4 Discrepancy analysis

In chapter 3 a study where variant translations were the objects in focus was presented. In this investigation the data used in the questionnaires were compiled more or less manually. The first step was to use a program to extract the recurrent sentences in both the source and target texts. Then the frequency data for the sentence types were compared and with these as a starting point the parallel text were examined in order to identify possible translation variants for identical source sentences. The process of extracting inconsistencies is straightforward, so we designed a simple tool, the Discrepancy Tool that would do the same work automatically, given a sentence-aligned text.

The task for the Discrepancy tool is primarily to identify the inconsistencies, in the following way:

1. Identify all repeated sentence types from the source text.

2. For each repeated sentence type, collect all target sentences.

3. Record whether the translations are consistent (i.e., the source sentence is translated with a single target sentence), or inconsistent (i.e., the source sentences are translated with two or more target sentences).

In conjunction with step 3, the Discrepancy tool will also produce numerical data and a set of measures that will clarify the nature of the discrepancies.

The input to the tool is a sentence aligned bitext. The records are tab separated and also contain a field for information about the sentence mapping relation, for example 1-1, 2-1, 1-2, 2-2, etc. The discrepancy analysis is performed on the actual records of the translation database. If a record holds more than one sentence, the analysis does not split up the record, instead the whole record is treated as the unit for discrepancy analysis. Optionally, all translation pairs that are not 1-1 sentence translations could be excluded from the discrepancy analysis. The default is to analyse the translation database from source to target, that is, to find inconsistent translations of repeated source sentences, but it is also possible to analyse it in the opposite direction, from target to source, which will identify a kind of "over-standardisation" on the part of the translators. In the last case, the tool identifies a list of source sentences that have been assigned the

same translation. Apart from the listings of the actual sentence types, there is also information on some general characteristics from the translation database. The example in Figure 13 illustrates the types of extracted data.

```
The information comes from the file D:\TEXTDB\XL5\xl5.alg,and has been treated as a source -> target translation.
Source sentence types:          All (types): 9957  All (instances): 12589
Consistently translated (types): 9785              Inconsistently translated (types): 172
Repeated source sentence types: All (types): 860   All (instances): 3492
Consistently translated (types): 688               Consistently translated (instances): 2336
Rtype= 8.64 repeated source types/all source types (x100)
Rinst= 27.74 repeated source instances/all source instances (x100)
ICtype= 20.00 inconsistent type translations/all repeated source types (x100)
ICinst= 33.10 inconsistent instance translations/all repeated source instances (x100)
ICtot-type= 1.73 inconsistent type translations/all source types (x100)
ICtot-inst= 9.18 inconsistent instance translations/all source instances (x100)
```

Figure 13. Sample of output from the discrepancy tool

The information provided concerns the number of sentence types and sentence instances, number of consistently translated sentence types and instances, number of repeated source sentence types and instances and how many of these that are consistent and inconsistent. At the end of the output above, percentages for various kinds of data are given. *Rtype* describes the percentage of all repeated sentence types compared to all source sentence types. *Rinst* shows the same but for instances instead of types. *ICtype* and *ICinst* describe the percentage of inconsistent sentence translations in relation to the set of all repeated source sentences. If the translation was totally consistent, both *ICtype* and *ICinst* would be zero. And finally, *ICtot-type/inst* gives us information on the proportion of inconsistencies in relation to all source sentences.

Table 38 shows an example of a source sentence that has been translated inconsistently and therefore has been identified by the discrepancy tool.

Table 38. Repeated source sentence with four different translations

	Sentence	**FRQ**	**Position & mapping relation**
SOURCE:	**Follow the directions in the Wizard dialog boxes.**	5	
TARGET 1:	Följ instruktionerna som visas i guidens dialogrutor.	1	7886 MAP 1-1
TARGET 2:	Följ instruktionerna i guidens dialogrutor.	2	8183 MAP 1-1, 8262 MAP 1-1
TARGET 3:	Följ anvisningarna i dialogrutorna.	1	10924 MAP 1-1
TARGET 4:	Följ instruktionerna i dialogrutorna som visas i guiden.	1	15051 MAP 1-1

Here the sentence "Follow the directions in the Wizard dialog boxes." occurs five times in the source text, but has been translated in four different ways, where the FRQ column for TARGET 2 indicates that one of the alternatives has been used twice.

If we take the same parallel text and reverse the analysis, that is, from target to source, we get data on potential synonym source sentences. Two examples from such an analysis are shown in Table 39.

Table 39. Two target sentence types with three synonym sources each

	Sentence	FRQ	Position & mapping relation
TARGET:	**Mer information finns i avsnittet "Ange relationer mellan tabeller" i kapitel 7, "Grunder för tabeller".**		
SOURCE 1:	For more information, see "Setting Relationships Between Tables" in Chapter 7, "Table Basics."	1	695 MAP 1-1
SOURCE 2:	For details, see "Creating Relationships Between Tables" in Chapter 7, "Table Basics."	1	5330 MAP 1-1
SOURCE 3:	For more details, see "Setting Relationships Between Tables" in Chapter 7, "Table Basics."	1	8943 MAP 1-1
TARGET:	**Välj Kör på Fråga-menyn eller klicka på Kör i verktygsfältet.**		
SOURCE 1:	Choose Run from the Query menu, or click the Run button on the toolbar.	1	6180 MAP 1-1
SOURCE 2:	From the Query menu, choose Run (or click the Run button on the toolbar).	1	6232 MAP 1-1
SOURCE 3:	Choose Run from the Query menu (or click the Run button on the toolbar).	2	6416 MAP 1-1, 6493 MAP 1-1

What the data in Table 39 reveal is that the translators have standardised the translations of variant source sentences. The implications for such data are that there is a possibility that the source text could become more standardised and that the source writers (or copy editors) could use information on synonym source sentences if they produce a new version of the text in question.

In principle, one could say that the discrepancy tool identifies possible synonym sentences on both the target and source side of a parallel text, because of the recurrent nature of the corresponding side.

In chapter 6 the application of the discrepancy tool on a number of parallel texts will be described.

5.5 Bilingual concordancing

The last component of the DAVE package is the Bilingual concordance program. As described previously in section 2.6.3, bilingual concordance programs can be regarded as a data acquisition tool, but can also be regarded as an aid when you

want to examine or verify different aspects of translations, for example the use of terminology. The Bilingual concordance component included in DAVE can be run both as a batch program where several searches can be specified beforehand or as an interactive tool. The user interface is shown in Figure 14 below.

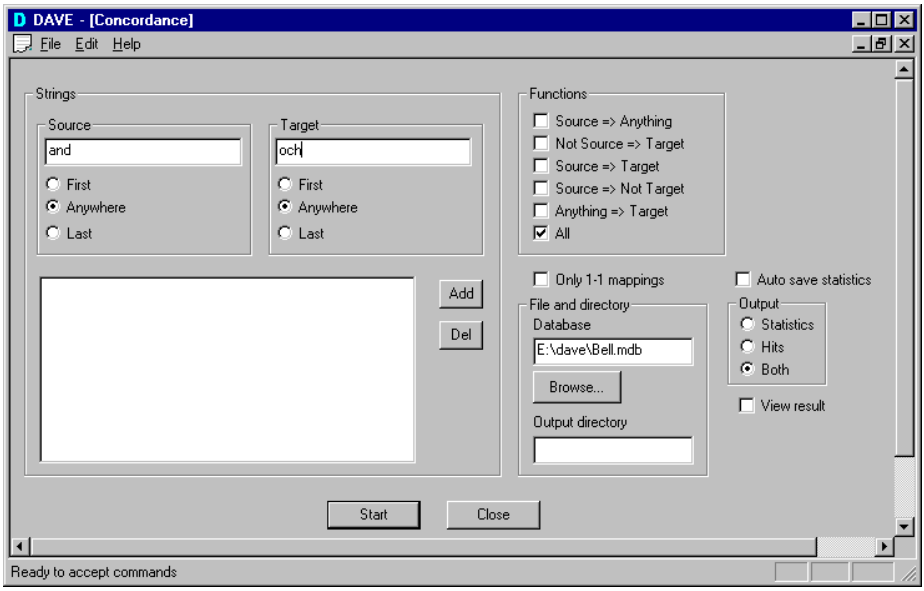


Figure 14. Bilingual concordance component

There are five different logical ways the parallel text can be searched for, indicated by the options under "Functions" above:

1. *Source =>Anything*, which means that a given source unit will be looked up and all the sentence pairs where it occurs will be presented.

2. *Not Source =>Target*, which means that the sentence pairs containing the specified target unit will be presented when they do no co-occur with the specified source unit.

3. *Source =>Target*, which means that all sentence pairs containing both the specified source unit and the specified target unit will be presented.

4. *Source =>Not Target*, which means that all sentence pairs containing the specified source unit but not the specified target unit will be presented.

5. *Anything =>Target*, which means that a given target unit will be looked up and all the sentence pairs where it occurs will be presented.

Apart from the different search methods, the user can specify whether the search units should start or end a record (the default is that the units could be placed anywhere in the sentences). It is also possible to specify multiple searches and

have them run in one batch. In this case the results will be written to different files.

The results from a search action in the bilingual concordance component include both the actual sentence pairs and also a brief description of the number of hits. If the search pair was given as "prevent – hindra*"[12], and all the five presentation functions are used, the resulting figures could be presented as in Figure 15 below.

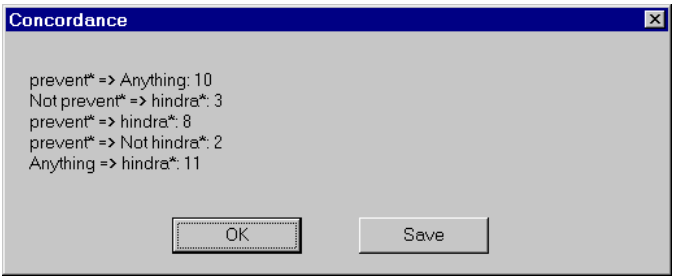


Figure 15. Numerical data from a bilingual concordance search

The data reveal that there are 10 instances of "prevent" and 11 of "hindra". Out of these "hindra" occurs three times when "prevent" is not present in the source sentence, and, likewise, there are two occurrences of "prevent" which do not co-occur with "hindra".

The actual occurrences of the hits are also presented in a separate window (or written to a file). The two instances where "prevent" is not translated by "hindra" are shown below:

> SOURCE: The crucial issue would be the guarantees of military security and the **prevention** of terrorism.
> TARGET: Själva kärnfrågan skulle bli garantier för militär säkerhet och skydd mot terrorism.
>
> SOURCE: An international organization that includes the International Atomic Energy Agency must be set up, with power to **prevent** the introduction of nuclear weapons into the Middle East."
> TARGET: En internationell organisation som också omfattar Internationella atomenergikommissionen måste upprättas, med befogenheter att förhindra införandet av kärnvapen i Mellersta Östern."

The use and application of the bilingual concordance program will be illustrated in Chapter 6.

[12] The asterisk used in the example is a wildcard, used to capture instances of *hindrar, hindrade, hindras, hindrat, hindrats*, etc. The database engine used for the bilingual concordance component is Microsoft Access, which therefore makes it possible to use all the regular expression possibilities included in that package

5.6 Summary

The tools presented in this chapter serve as an illustration of various tools that can be used to diagnose source and target texts, build parallel texts, extract monolingual and bilingual data as well as evaluate translations. The tools serve a double purpose; from one perspective they can be used as a way for translators to make their translations more efficient, and from another perspective they can assist the translation scholar in creating parallel texts and also support various means of accessing and studying existing translations.

This chapter also serves the purpose of presenting the tools that were used to build and analyse the Linköping Translation Corpus (LTC). The result of this analysis is the subject of the next chapter of the thesis.

6 Analysis of the translation corpus

In the chapter 4, the parallel texts under study have been presented and after independent analyses of the source texts and target texts respectively, a number of general observations regarding the characteristics, and, to a certain extent, the relationship between the source text and the corresponding target text were also shown. However, to be able to get a deeper understanding of these relationships, there is a need to look at the parallel text more closely. In particular, the focus is on whether there are any observable differences in the translations as far as text type and method of translation are concerned, especially for the distinctions of source- vs. target-orientation, consistency and correspondence. The following types of extracted data from the Linköping Translation Corpus (LTC) are presented:

- Sentence mappings (extracted from the MAP field of the sentence-aligned texts)
- Consistency and variation (using the Discrepancy tool)
- Correspondences between a sample of lexical items (using the Bilingual Concordance component).

6.1 Sentence mappings

When the translation corpus was created, the number of sentences included on each side of an alignment pair was added to the translation corpus.

One hypothesis is that translation memories contribute to more translations of the 1-to-1 type (*one* source sentence-*one* target sentence), because present TM systems are designed to handle this type of correspondence. Specifically, one would assume that sentences present in the translation memory would have a higher likelihood of being translated 1-to-1.

The default, and definitely most frequent, mapping in LTC is 1-1, that is, one source sentence is translated by one target sentence. By analysing the MAP field of the translation corpus we can collect information about the number of different sentence mappings that each translation holds. In Table 40 sentence mapping data are shown for all the translations in the corpus except ATIS. (The ATIS text contains only 1-1 mappings and is therefore not really interesting.)

Table 40. Sentence mappings from the parallel texts (excluding ATIS)

	Access		**Excel**		**OS2**		**InfoWin**	
	No.	*%*	**No.**	*%*	**No.**	*%*	**No.**	*%*
Pairs	14704		12589		11932		7771	
1-1	14169	*96.36*	12107	*96.17*	10444	*87.53*	7519	*96.76*
1-0	21	*0.14*	15	*0.12*	408	*3.42*	107	*1.38*
0-1	4	*0.03*	6	*0.05*	253	*2.12*	7	*0.09*
2-1	121	*0.82*	26	*0.21*	390	*3.27*	66	*0.85*
1-2	376	*2.56*	426	*3.38*	308	*2.58*	70	*0.90*
2-2	2	*0.01*	0	*0.00*	25	*0.21*	0	*0.00*
2-3	0	*0.00*	1	*0.01*	5	*0.04*	0	*0.00*
3-2	1	*0.01*	0	*0.00*	18	*0.15*	1	*0.01*
1-3	7	*0.05*	6	*0.05*	25	*0.21*	0	*0.00*
3-1	2	*0.01*	0	*0.00*	40	*0.34*	1	*0.01*
Rest	1	*0.01*	2	*0.02*	16	*0.13*	0	*0.00*

	Client		**Gord**		**Bellow**	
	No.	*%*	**No.**	*%*	**No.**	*%*
Pairs	2426		12254		4209	
1-1	2386	*98.35*	11112	*90.68*	4122	*97.93*
1-0	4	*0.16*	4	*0.03*	2	*0.05*
0-1	0	*0.00*	0	*0.00*	0	*0.00*
2-1	1	*0.04*	41	*0.33*	6	*0.14*
1-2	35	*1.44*	999	*8.15*	77	*1.83*
2-2	0	*0.00*	17	*0.14*	0	*0.00*
2-3	0	*0.00*	3	*0.02*	0	*0.00*
3-2	0	*0.00*	1	*0.01*	0	*0.00*
1-3	0	*0.00*	66	*0.54*	2	*0.05*
3-1	0	*0.00*	1	*0.01*	0	*0.00*
Rest	0	*0.00*	10	*0.08*	0	*0.00*

As can be seen from the table, the most common sentence mapping is 1-1 for all texts, which is hardly surprising. But if we look at the relative distribution of the different mappings, we see that most of the texts have 1-1 mappings within the interval of 96.36 to 98.35 per cent, and that the second most used sentence mapping is 1-2 followed by 2-1. The two texts that deviate from the others are the OS2 text and the novel by Gordimer.

The OS2 text has a strikingly high proportion of deletions (1-0) and insertions (0-1) which indicate that the translation is not particularly close to the original, but is rather a kind of communicative, more target-oriented translation (cf. Newmark (1988) and the discussion on page 17). However, the translation of this text has been made with the aid of a translation memory tool, which contradicts a target-oriented translation as translation memories should actually steer translators towards source-oriented translation. For the time being we can only note that there seems to be something strange about the OS2 translation, given the use of translation tools and the fact that data from the sentence mappings give us another message. But we will return to this observation later.

The second translation that differs from the others is the Gordimer text. Here there is only one aspect that seems peculiar, and that is the relatively high proportion of 1-2 mappings. Over 8 per cent of the pairs are instances of when one English sentence has been translated with two Swedish sentences. The explanation for this has to do with at least two different uses of punctuation characters in English and Swedish. First, in English the semicolon is used more often than in Swedish as a delimiter between main clauses, which means that perhaps we should have classified the semicolon as a possible sentence delimiter during the alignment process. Secondly, in English a comma may precede an utterance (within quotation characters) whereas in Swedish the line will commonly be indicated with a colon. The two different uses of semicolons and commas are shown in the two examples from the Gordimer novel in Table 29. The actual positions that were discussed above are underlined in the source and target texts.

Table 41. Different uses of the semicolon and the comma in English and Swedish

Source	Target	Map
It was she who had given her glass to him that night at the Independence party; the Pole who had danced the gazatska became the man with whom he gravitated to a quiet comer so that they could talk about the curious grammar-structure of Gala and the Lambala group of languages.	Det var hon som hade låtit honom överta sitt glas under självständighetsfesten. Polacken, som hade dansat en gazatska, blev den han så småningom drog sig undan med till ett lugnt hörn så att de kunde tala om den egendomliga grammatiska uppbyggnaden i gala- och lambalaspråken.	MAP 1-2
A youthful black official at passport control said uncertainly, "Just a minute.	En ungdomlig, svart tjänsteman i passkontrollen svarade litet osäkert: "Ett ögonblick.	MAP 1-2

If we regard semicolons as sentence delimiters as well as commas in sequences of *<comma-space-quotation mark-uppercase letter>*, and then recalculate the proportions, it turns out that the proportion of 1-1 mappings increases to 95.94 per cent, which takes the Gordimer text up to roughly the same relative proportions as the other texts (except the OS2 text). A flexible method to handle "unusual" sentence boundaries has been suggested by Palmer and Hearst (1994) which may help to improve sentence alignment.

Finally, a comment on the ATIS text. All the sentence mappings are 1-1 in this text. This is a hard coded feature of the program. Furthermore, the ATIS domain does not require 1-2 and or 1-0 mappings.

6.1.1 Comments on sentence mappings

The majority of the translations in the Linköping translation corpus contain 1-1 sentence mappings to the degree of a 96-98.35 per cent interval. One of the translation memory transaltions (OS/2) has a significantly lower proportion of 1-1 mappings, 87.53 per cent, which can indicate a more communicative kind of translation i.e., less source oriented than the others. The Gordimer text showed

that the sentence segmentation could cause problems if the standard way of only using periods, question marks and exclamation marks are regarded as sentence delimiters. However, if semicolons were included in the set and a special case was made for commas used in reported speech, then the proportion of 1-1 mappings could be seen as being on par with the other translations.

6.2 Consistency and variation revisited

In Chapter 3 an empirical investigation into translators' attitudes towards consistency and variation was presented. The study showed that translators prefer consistent translations, which makes translation memory-based tools look like the desired tool. However, it was also found that translators may have difficulty in accepting a previous translation from the translation memory, which may cause frustration among the translators when they are more or less forced to accept a given suggestion.

One solution to this problem is to use the discrepancy tool that is applied after all text has been translated, at the revising stage.[13] This tool would highlight all inconsistencies in the translation database and make it possible for the editor to revise the translations after the initial translation has been completed. The translators would not have to be totally restricted to previous translations; instead the consistency checking could be postponed to the revising stage. Obviously, the discrepancy tool does not have to be used in conjunction with a translation memory tool, it could just as well be applied to a fully manual translation provided that the translation database is created with some kind of sentence alignment program. Such programs exist both as commercial software (Trados' Talign and WinAlign as well as IBM's Visual Align) and in many variants from the academic world often originating from the algorithm presented by Gale and Church (1991).

Another cause of frustration among translators is that the source text is badly edited or just of poor quality. In our questionnaire many respondents expressed this as a major obstacle for producing high quality translations. In many cases, the translators actually detect unnecessary variation in the source text and "improve" the text by making these variants consistent. Given a translation database (a translation memory or a sentence aligned source and target text), it would be just as easy to produce a list of all "inconsistent source sentences", that would contain all the instances where the translators have identified synonym source sentences and translated these uniformly. This would constitute useful feedback to the technical writers who produced the source text, who could use this information for the next version of the source text.

Any translation of a repetitive source text can be placed somewhere between the endpoints of a consistency continuum, between maximal *consistency* and maximal *inconsistency*. In a maximally consistent translation, all repeated source sentence types have been translated uniformly and in a maximally inconsistent translation, all source sentences have different translations. As we will see in the

[13] Results from discrepancy analysis have previously been reported in Merkel (1996).

rest of this chapter, maximal consistency cannot, and probably should not, be found in real translations, even if they are translated with the aid of translation memory software. There are no empirical foundations for stating any preferred percentages of degrees of consistency, so the different figures provided in the study are merely used as a starting point for measuring and discussing consistency.

6.2.1 Discrepancy analysis of six translations

The discrepancy tool was applied to the eight different bitexts which form the Linköping Translation Corpus. The two novels in LTC are not really interesting applications for consistency checking, partly because they are not at all as repetitious as the manuals and partly because consistency is not something necessarily aimed for in literary translation. Nevertheless, the data from the novels could prove to be interesting for translation studies. The MT-translated dialogues are not included here, as they contain no repetitions at all, and consequently there are no inconsistencies to be measured. Instead we will concentrate on the software manuals. Let us begin by looking at the data from the manually translated software manuals, Excel User's Guide and Access User's Guide from Microsoft (Table 42). The table below shows the output from the Discrepancy tool (described in section 5.4). I repeat the explanations for the discrepancy measures here: *Rtype* describes the percentage of all repeated sentence types in relation to all source sentence types. *Rinst* shows the same but for instances instead of types. *ICtype* and *ICinst* describe the percentage of inconsistent sentence translations in relation to the set of all repeated source sentences. If the translation was totally consistent, both *ICtype* and *ICinst* would be zero. And finally, *ICtot-type/inst* gives us information on the proportion of inconsistencies in relation to all source sentences.

Table 42. Discrepancy data for two manually translated software manuals

	Access		Excel	
CATEGORY	**S->T**	**T->S**	**S->T**	**T->S**
Sentence types	10849	10970	9957	9950
Sentence instances	14704	14704	12589	12589
Consistent types	10502	10771	9783	9586
Inconsistent types	347	199	174	194
Repeated types	1272	1111	860	846
Repeated instances	5127	4845	3492	3685
Consistent repeated types	925	912	686	652
Consistent repeated instances	3384	3290	2332	2357
Rtype	11.72	10.13	8.64	8.68
Rinst	34.87	32.95	27.74	29.27
ICtype	27.29	17.91	20.23	22.93
ICinst	34.00	32.09	33.22	36.04
ICtot-type	3.20	1.81	1,75	1,99
ICtot-inst	11.85	10.58	9.21	10.55

The data in Table 42 show the discrepancies in both directions, source to target, and target to source, indicated in the columns S->T, T->S respectively. The

texts are roughly of the same size, Access being slightly more repetitious than Excel. When we look at the consistency measures, we find that the Access translation shows more signs of being inconsistent (*ICtype*: 27.29 vs. 20.23 per cent). The Excel translators have also identified more inconsistencies in the source than the Access translators have (*ICtype*: 22.93 vs. 17.91), but these are only small differences. What is clear is that both translation teams have "missed" a certain number of sentences that could have been translated consistently, but they have also "improved" the translation in relation to the source text by making synonymous source sentences have the same translation.

The Microsoft manuals are far from being 100 per cent consistent; if that had been the case, we would have had zero values for *ICtype*, *ICinst*, *ICtot-type* and *ICtot-inst*, both from source to target, and from target to source. For manual translations, however, this is what one could expect. Several translators working simultaneously on different parts of the documents must face problems in being consistent if they do not have the support of a translation memory. What is remarkable, in spite of the lack of computerised support, is that the figures for inconsistencies from target to source are as high as they are, which means that the translators somehow have identified a large set of different source sentences that have been translated uniformly.

Let us now turn to two translations that have been done with the aid of translation memories and compare them with the manual translations. A reasonable expectation here would be that the figure for *ICtype/inst* would drop to close to zero, approaching maximal consistency when applied from source to target, and that there will be high figures for *ICtype/inst* when applied from target to source. This last expectation depends highly on the use and success of the fuzzy matching techniques present in the translation memory tool. With fuzzy matching the translators ought to have good opportunities to detect minor differences in the source text and thus make the translation more uniform. The translations were made with the aid of IBM's Translation Manager/2 and stem from the period when this tool was introduced into the translation process at IBM Sweden.

Table 43. Discrepancy data for OS/2 Installation Guide and OS/2 User's Guide (TM tool)

	OS/2 IG		OS/2 UG	
CATEGORY	**S->T**	**T->S**	**S->T**	**T->S**
Sentence types	2615	2640	7054	7148
Sentence instances	3057	3057	8876	8876
Consistent types	2521	2567	6774	6914
Inconsistent types	95	73	280	234
Repeated types	190	196	685	638
Repeated instances	631	613	2507	2366
Consistent repeated types	95	123	405	404
Consistent repeated instances	252	324	1086	1240
Rtype	7.26	7.42	9.71	8.93
Rinst	20.64	20.05	28.24	26.66
ICtype	50.00	37.56	40.88	36.68
ICinst	60.06	47.15	56.68	47.59
ICtot-type	6.63	2.77	3.97	3.27
ICtot-inst	12.40	9.45	16.01	12.69

The two texts are considerably shorter than the Microsoft manuals, and they are slightly less repetitious if we measure the percentage of repeated sentences. Nevertheless, there are some strange things going on here. If we look at the *ICtype* figure (measuring the number of inconsistently translated sentences in relation to the set of repeated sentences), we see that the OS/2 translations are much more inconsistent than the Microsoft translations, in spite of the use of Translation Manager/2. In the *OS/2 Installation Guide*, fifty percent of all repetitions show signs of inconsistencies. In the *User's Guide* the corresponding figure is lower, but still higher than in the Microsoft translations, which were done without a translation memory tool. The ICtype figures in the reverse direction indicate, however, that the translators have found many inconsistencies in the source text and standardised this variation in the translation. The data in Table 43 seems to indicate that the use of the translation tool had very little effect on the translation as far as increasing consistency from source to target text. Furthermore, it raises questions on whether translation memory software really saves time and money, as is often promised by the producers of such software (cf. Schäler (1994)).

From the data alone it was difficult to figure out what had really happened during the translation project so I had to go back to the translators at IBM. In an interview with two of the translators who had been working on the company's translations before and after the introduction of TM/2, the following explanations were given:

At the beginning of the nineties IBM had had a long history of "rewriting". The source text was merely seen as some kind of guidance to what the translator/writer should produce. The goal was a coherent Swedish text, clearer and more concise than the original. What they aimed for was something that in Newmark's terminology was something of a "free translation" with a strong target language emphasis (Newmark 1988). In the translation guidelines for IBM, it is stated that the translator should review the source text critically,

delete unnecessary passages or repetitions, rearrange and regroup the source text before starting the translation (Ström and Windfeldt 1991). The translation culture at IBM at the beginning of the nineties was not exactly suitable for a swift adaptation to translation memory-based software. They had a translation team that was highly skilled in producing high-quality Swedish versions of American originals in a creative way. Some translators had an openly negative attitude towards the new tools, some people left and some protested in other ways, for example by ignoring TM/2 and sticking to the old ways of doing things. Initially there were also problems with the administration and distribution of translation memories, which also contributed to resistance from the translation teams.

The basic explanation to the strange discrepancy figures in the OS/2 translations according to my interviewees was that there had been a clash between new technology and an old translation culture, where the old culture was initially stronger than new tools.

Instead I looked at some other texts where they thought that the use of TM/2 had been more successful. These texts are IBM's documentation on InfoWindows and Client Access for Windows, where the translation of the former actually stems from about the same time as the OS/2 translations (1993), but the Client Access translation was done during 1995. A summary of the discrepancy data for these texts looks like the following:

Table 44. Discrepancy data for InfoWindows and Client

	InfoWindows		Client Access	
CATEGORY	**S->T**	**T->S**	**S->T**	**T->S**
Sentence types	5157	4881	1689	1658
Sentence instances	7771	7771	2054	2054
Consistent types	5121	4443	1663	1617
Inconsistent types	136	438	26	41
Repeated types	1076	1214	146	153
Repeated instances	3590	4104	882	920
Consistent repeated types	940	776	120	112
Consistent repeated instances	2974	2213	718	687
Rtype	20.47	24.87	8.64	9.26
Rinst	46.20	52.81	36.37	37.94
ICtype	12.64	36.08	17.81	26.80
ICinst	17.16	46.08	18.59	25.33
ICtot-type	2.59	8.97	1.54	2.47
ICtot-inst	7.93	24.33	6.76	9.61

Here the degree of inconsistencies is much lower from source to target, 11.64 and 17.81 per cent (*ICtype*), compared with the other IBM translations. The translators working for IBM said during the interview that they had seen significant changes in IBM's translation culture in Sweden. Now there was more pressure to reuse old translations, and adjust the translation to this recycling process. One example of this is that the more recent translation of Client Access contains 98.35 per cent 1-1 mappings (one source sentence - one target sentence)

whereas for the "older" OS/2 Installation Guide the corresponding figure was 81.91 per cent. This in itself is a good indication of the changes that have taken place.

6.2.2 Other Applications of Discrepancy Analysis

Restricting variation in the source text is also used for MT systems that adopt a controlled language approach. If a company moves in the direction of automatic translation, a discrepancy analysis of their previous translations will provide useful information on how to design such a controlled language. For use of controlled languages in MT is described in for example Adrians and Macken (1995) and Mitamura and Nyberg (1995).

With the increasing availability of text corpora, researchers, teachers and students within corpus linguistics and translation studies have access to new ways of exploring theoretical and descriptive branches of their fields, see for example Baker (1993). One direct application of the discrepancy tool would be to study variation of equivalent idiomatic dialogue in different languages, based on translation corpora. As mentioned earlier, the novels included in the analysis do not contain a high degree of repeated sentences. But the small set of repetitive sentences that do occur is interesting, because, at least in one of the two novels, the repetitions are from the fictitious characters' dialogue, which provide an interesting contrast between idiomatic dialogue patterns in English and Swedish.

Table 45. Idiomatic variation in the dialogue from the Gordimer novel

	Sentence	FRQ		Sentence	FRQ
TARGET:	**"Naturligtvis inte.**		**SOURCE:**	**I know.**	
SOURCE 1:	Of course.	1	TARGET 1:	Ja, jag vet.	2
SOURCE 2:	Of course not.	1	TARGET 2:	Ja, jag vet det.	1
TARGET:	**"Hur då, menar du?"**		TARGET 3:	Jag vet det.	1
SOURCE 1:	"To what?"	1	TARGET 4:	Jag vet.	1
SOURCE 2:	"What do you mean?"	1			
TARGET:	**"Herre Gud!"**		**SOURCE:**	**Of course.**	
SOURCE 1:	"Oh my God."	1	TARGET 1:	Naturligtvis inte.	1
SOURCE 2:	"Good God."	1	TARGET 2:	Naturligtvis.	3

The examples in Table 45 show the potential for using the discrepancy tool for corpus and translation studies. Note, for example, the translations of the English "Of course" where the expected Swedish equivalent would be "Naturligtvis". However, the corpus reveals that the context sometimes requires a negated variation as well ("Naturligtvis inte."); the same phenomenon is present in the other direction ("Naturligtvis inte" vs. "Of course" and "Of course not". The examples are taken from only one novel, but if large translation corpora containing dialogue were created, this would provide a starting point for studying idiomatic correspondences which go far beyond the phraseology, style and conventions described in dictionaries and translation text books.

6.2.3 Comments on Translation Memories and Discrepancy Analysis

Translation memory tools are designed for translation projects of repetitive texts and should in principle help the translators to increase both speed and quality of the translation task. However, if these tools will prove to be efficient depends on the quality of the source text, the quality of previous translation memories, the attitude towards new technology among the translators using such tools as well as to what extent repeated source text segments can be transferred to corresponding target segments.

In this investigation I have shown that both manually translated texts and texts translated with the aid of translation memories are more or less inconsistent. For manually translated texts the variations are what can be expected, but the high degree of inconsistency found in IBM's early translation memory translated manuals was due to a clash between an established translation culture and new technology. Over time the use of translation memories seems to have been more successful, if efficiency is measured as the degree of consistency. O'Brien (1999) and Schäler (1994) have both expressed their criticism towards translation memories (see sections 2.4.2 and 3.4) and their fears seem to be have come true judging from this discrepancy analysis.

From one perspective, the discrepancy tool should be seen as a translation checker that can be activated in the postediting phase of a translation project. The tool will help the editor to pinpoint inconsistencies in the translation, edit them or leave them as they are in the published translation. The editor will furthermore have the possibility to verify the translations in a translation memory before it is archived for future use or remove inconsistencies in an existing bank of many translation memories. This is important when the number of translation memories is steadily growing. If translation memories are not verified at one stage or the other, the translator may be facing so many translation alternatives that browsing through these options will take as much time as translating the sentence without the memory.

The discrepancy tool could also be seen as a complement to translation memory software and one of several utilities that will enhance the quality of the translations. Ideally, discrepancy analysis should be integrated with the translation memory software in order to verify and clean up the memory before archiving. It will help to make the translations more consistent, and also to extract information about unnecessary variation in the source text that can be fed back to the technical writers. Applying the tool to parallel texts will give a clear indication of how efficient the use of translation memory tools has been. If the inconsistencies between source and target are too numerous, then this may reveal that there is a serious problem in the company's management and configuration of the translation project.

From another perspective, the discrepancy tool could be seen as a tool for doing studies of translation. By for example, investigating what is and what is not translated consistently, the translation scholar can make a number of interesting observations, for example, when studying dialogue patterns in fiction.

6.3 Investigating Word Co-occurrences with the Bilingual Concordance component

Co-occurrence data could be useful in several applications. In contrastive linguistics, it could form the basis for extracting exactly those sentence pairs that contain the word(s) that are of interest to the scholar. Altenberg (1998) has developed a measure, *mutual correspondence*, that aims to capture the degree of correspondence in parallel corpora between pairs of words in English-to-Swedish translations and Swedish–to-English translations jointly. For example, if the English word "however" is always translated into the Swedish "emellertid" and "emellertid" is always translated by "however" in English translations then the Mutual Correspondence (MC) between "however" and "emellertid" is 100 per cent.

Altenbergs formula for calculating MC looks as follows:

$$MC = \frac{(A_t + B_t) \times 100}{A_s + B_s}$$

where A_t and B_t are the number of target items and A_s and B_s are the number of source items that are being compared in two translation directions. As the Linköping translation corpus only contains translations in one direction, namely into Swedish, mutual correspondence cannot be calculated, but it is possible to investigate the relative word co-occurrence rate (WCR) for a source and a target word, as follows:

$$WCR = \frac{2 \times (cooccur(A,B) \times 100}{freq(A) + freq(B)}$$

If a word A occurs 10 times in the source text, a target word B occurs 15 times in the target text and A and B co-occur 8 times, the WCR for A and B is 64 per cent (16/25). The measure actually takes into account the number of times a token of one of the words occurs in a co-occurrence relation in the corpus.

The MC measure will capture the extent to which two words are mutual translations of each other, while the WCR measure will measure the proportion of co-occurrence between one source and one target word given a corpus containing only one translation direction.

Altenberg measures the *translation bias* as the ratio of how many target tokens that are realised from the source word. For example, if the conjunction "and" occurs 207 times in the source text and it co-occurs with the Swedish conjunction "och" 164 times, then the *translation bias* for choosing "och" as the translation of "and" is 79 per cent (164/207).

The formula for translation bias (TB) can be expressed as follows:

$$TB(A,B) = \frac{cooccur(B,A)}{freq(A)}$$

which means a simple ratio between the number of times a target item, B, co-occurs with the source item, A, in relation to the total number of source items A.

Different hypotheses could be tested by investigating co-occurrence data and translation bias; for example, is it possible to conclude how source-oriented a target text is given only co-occurrence data of word pairs from different translations? Text-type specific translation corpora could provide information about what the standard co-occurrence rates would be for a core of word pairs. These pairs and their co-occurrence rates could then be tested on translations from the same text type and perhaps give an indication of how source-oriented the translations are.

Furthermore, co-occurrence rates could be used to investigate what word pairs that are most suitable to use as anchor words (Johansson and Hofland 1994), cognates (Simard et al. 1992) or cue words Wu (1994) in hybrid approaches to sentence alignment (cf. section 2.5).

The use of the bilingual concordance component from DAVE and the application of word co-occurrence rates to the corpus can be illustrated with a sample of word pairs from the Linköping translation corpus. The word pairs belong to four different categories: conjunctions, subjunctions, numerals and proper names/technical terms. In order to avoid too many figures, the data have been taken from one of the Microsoft manuals, namely Microsoft Access User's Guide (human translation of manuals), two of the IBM manuals (translation memory-based translations of computer manuals from two periods of time) and the Gordimer novel. It may be redundant to mention that the ATIS text (automatic translation) exhibit more or less 100 per cent word co-occurrence rates, which makes it unnecessary to present data from that text.

6.3.1 Conjunctions (But & And)

The English connector "but" is an interesting candidate for anchor words. In studies of English-French translation corpora, Salkie (1997) reports a 72-85 per cent correspondence between the English "but" and the French "mais". Altenberg (1998) refers to studies of English-Norwegian where Haaland (1997) reports high degrees of correspondence between the English "but" and the Norwegian "men" in the English-Norwegian Parallel Corpus (92-95 per cent). Altenberg measured the mutual correspondences between "but" and "men" to 85-90 per cent in the English-Swedish Parallel Corpus from Lund.

The distribution, word co-occurrence rate (WCR) and translation bias (TB) for "but" and "men" in the texts of the Linköping Translation Corpus are shown in Table 46 below.

Table 46. Distribution, co-occurrence and translation bias for "but" and "men"

	Microsoft Access	IBM OS/2	IBM Client	Gordimer
but-men	140	24	13	916
but-NOT men	36	37	2	122
NOT but-men	73	31	3	332
but-	176	61	15	1038
-men	213	55	16	1248
WCR	**71.98**	**41.38**	**83.87**	**80.14**
TB	**79.55**	**39.34**	**86.67**	**88.25**

None of the texts show co-occurrence values as high as the correspondence rates reported in the Altenberg study. However, the important point here is the relative differences between the texts, especially between the two IBM texts. The OS/2 translation have less than half the translation bias compared to the later Client translation, which seems to support the view that the Client translation is more source-oriented, and more adapted to the requirements of translation memory-based translation than the first attempt to use the technology (see also the discussion in section 6.2.1).

The other conjunction pair that illustrates the use of the bilingual concordance program is the pair "and-och", shown in Table 47 below.

Table 47. Distribution and co-occurrence for "and" and "och"

	Microsoft Access	IBM OS/2	IBM Client	Gordimer
and-och	2600	821	163	3294
and-NOT och	288	363	44	216
NOT and-och	434	422	38	783
and-	2888	1184	207	3510
-och	3034	1243	201	4077
WCR	**87.81**	**67.66**	**79.90**	**86.83**
TB	**85.70**	**69.34**	**78.74**	**93.85**

Here the differences between the two IBM translations are not as substantial as for "but" and "men", but the IBM Client translation can be judged as slightly more source-oriented than the OS/2 translation as indicated by the WCR and TB values.

6.3.2 Subjunctions (If & When)

The subjunction "if" in English has one major candidate for correspondence in Swedish, namely "om". The Swedish counterpart has however other usages than the strict conditional one. "Om" can, for instance, also be used as a preposition (*om fyra veckor – in four weeks* and *ha kunskaper om – have knowledge about*) as well as a particle (*göra om – redo*). In the string-based approach of DAVE, it is not possible to tell the difference between the different Swedish usages of "om", but it

is still possible to measure to what extent "if" co-occurs with "om". In Table 48 the distribution of this pair is shown.

Table 48. Distribution and co-occurrence for "if" and "om"

	Microsoft Access	**IBM OS/2**	**IBM Client**	**Gordimer**
if-om	1103	428	85	611
if-NOT om	89	156	10	41
NOT if-om	1152	508	128	1199
if-	1202	584	95	652
-om	2255	936	213	1810
WCR	**63.81**	**56.32**	**55.19**	**49.63**
TB	**91.76**	**73.29**	**89.47**	**93.71**

Although the co-occurrence rates are lower than in the previous word pairs, they are probably still high enough to qualify as good candidates for anchor words. The fact that the Swedish "om" has different usages is reflected in the relatively high frequency of "om" in Swedish compared to the frequency of the English "if". There are between two and three times as many "om" tokens as there are "if" tokens in the texts. If we only look at co-occurrences from one direction, namely English-to-Swedish, we see that the bias for choosing "om" as a translation for "if" is around 89 per cent for the IBM Client text (85 out of 95) whereas the same figure for the OS/2 translation is 73 per cent (428 out of 584). Apparently, most of the "om" tokens in the Swedish translations correspond to something other than "if", but when "if" occurs in the source text, between 73 and 93 per cent of these tokens are translated with the Swedish "om".

In Table 49 data about the pair "when-när" are shown.

Table 49. Distribution and co-occurrence for "when" and "när"

	Microsoft Access	**IBM OS/2**	**IBM Client**	**Gordimer**
when-när	764	164	39	368
when-NOT när	130	124	15	113
NOT when-när	556	207	36	353
when-	894	288	54	481
-när	1320	371	75	721
WCR	**69.02**	**49.77**	**60.47**	**61.23**
TB	**85.46**	**56.94**	**72.22**	**76.51**

The WCR values range from around 50 per cent for the IBM OS/2 translation to 69 per cent for the translation of Microsoft Access. Both the WCR and TB values for the two IBM texts support the tendency towards an increasing source-orientation.

6.3.3 Numerals (1)

Using similar looking strings (or tokens) as cognates is a common strategy in alignment systems, both for word alignment (Melamed 1995, 1999) and for sentence alignment (Johansson and Hofland 1994). The fact that most numbers occur in identical form in both the source and the target is something that can be utilised in alignment. To illustrate this phenomenon, data for the most frequent numeral in the corpus, namely the number "1", is shown in Table 50.

Table 50. Distribution and co-occurrence for "1" and "1"

	Microsoft Access	**IBM OS/2**	**IBM Client**	**Gordimer**
1-1	888	647	119	5
1-NOT 1	57	124	1	3
NOT 1-1	158	63	2	7
1-	945	771	120	8
-1	1046	710	121	12
WCR	**89.20**	**87.31**	**98.76**	**50.00**
TB	**93.97**	**83.92**	**99.17**	**62.50**

The figures in the table speak for themselves; numerals seem to be an important category if one is looking for "safe" co-occurrences or anchor words. The only thing that needs spelling out here is that digits are sometimes spelled out in letters (for example, "1" is translated as "ett", Eng. "one") and this is the case for the majority of the tokens in the corpus where digits do not co-occur.

6.3.4 Proper names and terms

Usually proper names are translated with little or no changes and they are therefore also excellent candidates for using as anchor words in alignment. Furthermore, the fact that, at least in English and Swedish, most proper names are spelt with an initial upper-case letter makes them easier to identify in the text. A method for identifying and extracting proper names has been proposed by Danielsson and Mühlenbock (1998). As can be seen in Table 51, proper names from the Gordimer novel are more or less always transferred in more or less identical form:

Table 51. WCR and TB values for proper names in the Gordimer novel

Proper name	WCR	TB
Bray	99.87	99.35
Rebecca	99.46	100.00
Mweta	99.58	99.79
Shinza	99.41	99.61
Kalimo	100.00	100.00
Wentz	99.38	100.00
Africa (Afrika)	99.62	99.62
England	99.10	98.21
Europe (Europa)	93.10	87.10
London	100.00	100.00

As can be seen here, proper names co-occur almost one-to-one in the novel (the same types of values are, by the way, also present in the Bellow novel). The few times where there is no co-occurrence can in most cases be explained by the use of pronouns instead of the name, as in the following example:

SOURCE: When the child had gone she sat with her hands between her spread thighs, staring at the typewriter.
TARGET: När flickan hade gått satt Rebecca med händerna mellan sina utspärrade lår och stirrade på skrivmaskinen.

In the computer manuals, product names and system-specific terminology also show high WCR and TB values, but not quite as high as for proper names in the novels, see Table 52 below.

Table 52. Word co-occurrence rates and translation bias for computer terms in the computer manuals

Term	Microsoft Access		IBM OS/2		IBM Client	
	WCR	TB	WCR	TB	WCR	TB
command (kommando)	89.49	84.01	75.36	75.20	92.06	93.55
button (knapp)	64.29	48.65	83.97	75.29	92.38	90.65
menu (meny)	97.39	99.08	70.41	80.70	92.39	87.63
icon (ikon)	92.96	86.84	82.65	89.12	96.48	95.05

The IBM Client translation stands out as the most source-oriented of the manuals, especially in comparison with the OS/2 translation. The relatively low values for "button (knapp)" in the Microsoft Access translation are actually due to the tendency to omit certain classifiers in the Swedish translations. The English "Click the OK button" is commonly translated by "Klicka på OK". See also sections 11.4.4 and 11.4.6.

6.3.5 Comments on the use of bilingual concordancing

By using the bilingual concordance component from the Dave toolbox, it has been shown how one can investigate word co-occurrence in translation corpora in order to find suitable candidates for anchor words. The brief analysis presented here confirms the view that the best candidates for anchoring words can be found among cognates (numbers and proper names) an, to a certain extent, technical terms. Conjunctions and subjunctions can also function as potentially good candidates for most texts. Coupled with suitable alignment techniques, it is reasonable that empirical investigations into what the lower bound for co-occurrence or translation bias rates is.

The Bilingual concordance component is a useful tool for the contrastive linguist, the translation scholar and the language engineer. The contrastive linguist can compile statistical data on co-occurrence and extract sentence pairs from parallel corpora that are specifically interesting for a certain contrastive study. The translation scholar may be more inclined to study translation bias; that is, given a certain source object, what are the preferred choices of the translators as they appear in the text. The translation scholar will focus on the translation direction of the text, whereas the contrastive linguist will be more interested in the relationship between two language systems. The Bilingual concordance component will aid both types of linguists in their endeavours. Furthermore, for system developers of translation support software, investigations into resources like anchor words, co-occurrence relations, etc. are empirical foundations for producing better and more appropriate software for translators.

Given large translation corpora, preferably divided into text genres, it may be possible to perform further tests to see whether single word co-occurrences can be taken as indicators of how source- or target-oriented a certain translation is. This brief study revealed that the last of the three translation-memory translations of IBM texts showed a tendency to being closer to the source text than the first translation (IBM OS/2). This is in accordance with the earlier observations of discrepancy data where the consistency in the later translation had increased considerably.

6.4 Summary

In this chapter, three of the components in the DAVE package have been applied to the Linköping Translation Corpus in order to extract data that shed some light on the relationship between the respective source and target texts and the different text types. Information on sentence mappings reveal that five of the translations show 1-1 translations (1 source sentence – one target sentence) in the region of 96.36 to 98.35 per cent. Three of the translations deviated from this range for various reasons: the machine translated ATIS text (100 per cent), the OS/2 manual from IBM (87.53 per cent) and the Gordimer novel (90.68 per cent). All the sentence translations of the ATIS text are 1-1 as this is built into the MT system. The OS/2 translation was shown to contain a high proportion of deletions and additions (1-0 and 0-1) and the Gordimer novel exhibited a

deviation from the other texts as far as the use of punctuation is concerned (depending on semicolon and comma use in dialogue).

The discrepancy analysis of LTC revealed the apparent inefficient use of translation memories for the IBM translations. This was confirmed in interviews with the translators and explained by a combination of technical problems and a clash of culture with the previous way translations were done at the company. In fact, this observation confirms the problem brought to surface in chapter 3 where the translators were shown to have difficulties in accepting a suggested translation. The discrepancy analysis was also shown to be usable in studying other phenomena, such as variations in the dialogue included in novels.

Finally, the bilingual concordancing tool was applied to LTC and its usefulness was illustrated as a tool for the contrastive linguist, the translation scholar and the language engineer. Anchor words used for sentence alignment can be empirically tested with measures such as the Word Co-occurrence Ratio (WCR) and Translation Bias (TB).

Judging from the data presented in this chapter, the use of translation memories does not necessarily result in more source-oriented translations. The OS/2 translation contains too many non-1-1 sentence mappings to be judged as source-oriented. Furthermore, word co-occurrence and translation bias data show that the OS/2 translation is considerably less consistent than the rest of the manuals. When the manually translated Microsoft Access User's Guide is compared to the translation memory-translated IBM manuals, there are no clear tendencies. On the one hand, the Microsoft manual seems to be more source-oriented and consistent than the OS/2 translation, and on the other hand, less source-oriented and less consistent compared to the IBM Client translation. The most likely explanation to this observation is that the IBM OS/2 and the IBM Client translations were done at different stages at IBM. The OS/2 translation was done in 1993 when the technology was new to the translators. In 1995 when the Client manual was translated, translators had adapted their translation to the tool, which led to higher consistency and higher co-occurrence ratios.

Consistency in translation is important for technical translations, but is not a major issue for translation of fiction, mainly because of the lack of sentence recurrence in the novels. On the lexical level, the novels exhibit an almost 100 per cent word-co-occurrence for proper names.

Altogether, the tools complement each other in revealing and measuring interesting differences between translations, as well as in investigating translation norms.

7 Linköping Word Aligner (LWA)

In the last five to ten years the interest in systems that align (or link) words and phrases in a source text to corresponding target units has increased steadily. In this chapter and in the rest of the thesis, the notion of *word alignment systems* is used as a general term for systems that align linguistic units below the sentence level across two languages. These linguistic units could be expressed as single words, phrases, terms or collocations. The majority of the word alignment systems described in the literature fall into two main categories: (1) Full-text alignment systems, and (2) Bilingual lexicon extraction systems. Below these main categories, it is possible to make further divisions into, for example, bilingual concordancing and bilingual information retrieval for the first category, and technical terminology systems and systems that compile lexicons automatically for specific systems or specific uses for the second category.

A word alignment system can be seen as kind of data extraction tool applied to a bitext (see section 2.6.2). There are several potential applications for word alignment, both for research activities and for more applied purposes. Full-text alignment systems will actually pinpoint the correspondences between running words in the source text and words in target text. This would be a helpful complement to bilingual concordance systems (see also sections 2.6 and 5.5) as it would be possible to present both a source item and a corresponding target item for the user with linguistic context. Within natural language processing, word alignment systems have been used for different purposes, e.g. as a way to build up bilingual lexicons for machine translation systems (cf. Melamed 1998b).

Linköping Word Aligner (henceforth LWA) is a system for word alignment that has been in operation since the fall of 1997. In this chapter the LWA design and implementation are presented, as well as the basic setup for two different tasks involving two different language pairs, English/Swedish and French/English. In chapter 8 the problem of evaluating word alignment systems in general is discussed, and a proposal for a specific way of evaluation is made that includes software developed in Linköping - the PLUG Link Annotator and the Link Scorer. In chapter 9, the LWA system is evaluated in three different steps.[15]

LWA was developed to process the Linköping translation corpus described in chapter 4. The primary purpose of the system is to explore the possibilities of generating lexical data from parallel texts in the form of translation lexicons and bilingual concordances. At the time several results on word alignment had already been published (e.g. Brown et al. 1991, Kay and Röscheisen,1993), Smadja 1993, Kupiec 1993, van der Eijk 1993, Fung and Church 1994, Wu 1995,

[15] Chapters 7, 8 and 9 are partly built on Ahrenberg, Andersson and Merkel (1998a, 1999), Merkel, Andersson and Ahrenberg (1999), Merkel and Ahrenberg (1999) and Merkel (1999).

Chang and Ker 1996) Kaji and Aizone 1996, Macklovitch and Hannan 1996, Melamed 1995, 1996a, 1996b, 1996c, 1997a, 1997b, Fung and McKeown 1997, Kitamura and Matsumoto 1996, Tiedemann 1997, and Resnik and Melamed 1997) and it seemed fairly evident that a system that combined statistical and non-statistical "linguistic" information would be the best choice. This still left many options open, however, and it was much less clear what knowledge sources would actually be useful or available.

Moreover, previous work had focused on French/English texts, in particular the extensive Hansard Corpus, with some attempts also being made at parallel texts from structurally very different languages such as English/Japanese or English/Chinese. While English and Swedish are relatively similar in structure, there are some crucial differences that must be taken into account. In English a large number of lexical units and technical terms are multi-word compounds, while the corresponding units in Swedish and other Germanic languages are often single-word compounds. Two examples are English *file manager* corresponding to Swedish *filhanterare*, and English *sewage disposal plant* corresponding to Swedish *vattenreningsverk*. Thus, a desideratum for LWA was that it should be able to handle multi-word units and single-word units alike (cf. Jones and Alexa 1997, Blank 1999).

Other important requirements for LWA were that it should be *modular*, so that different information sources could be combined in different ways, and easily configurable for different (European) language pairs. This latter requirement made us settle for a *knowledge-lite* approach, i.e. a system whose modules can be provided with the necessary language data quickly and inexpensively. For this reason "knowledge-intensive" sources were excluded from LWA, such as machine-readable bilingual dictionaries, part-of-speech taggers and lemmatizers. It is true that such resources are becoming more readily available and future versions of LWA may well make use of them. Even though it is likely that knowledge-intensive sources will give better performance, knowledge-lite systems still have an interest. Apart from portability they are likely to give sufficiently good results for many purposes, e.g. if the output is to be used by a human user for the creation of a complete word-aligned bitext. LWA actually provides two kinds of output: a set of links that provide the basis for a bilingual concordance, and a list of link types from the text that constitute a partial translation lexicon for the bitext.

In the rest of this chapter the overall design and implementation of the system is described as well as its basic setup to handle English/Swedish alignments. Then a presentation of the adaption of LWA to French/English texts for the Arcade word alignment contest (Véronis and Langlais 1999) follows.

7.1 The system

The objective for LWA is to find word links in a bitext and generate a non-probabilistic translation lexicon from it. It is assumed that the bitext is already correctly aligned at the sentence or paragraph level. These aligned sentences or paragraphs are referred to as *bitext segments*.

In this section an overview of the assumptions behind the LWA system are presented as regards the objects and relations found in bitext segments, and how they are implemented in the LWA system. Then a description of how the system is used is given as well as further details on its mode of operation.

7.1.1 Underlying assumptions

A given bitext segment can be analysed as a *sequence of tokens and delimiters*. Most characters are unique to either tokens or delimiters, but some characters that may occur either as delimiters or word-internally, e.g. the full stop, are treated by rules. Thus, the tokenizer can deal with abbreviations with internal punctuation marks. It also regularises clitics to tokens (e.g. *can't* is regularised to *can not*) and capital letters are made lower-case.

A segment is also viewed as a *partially ordered bag of lexical units*. It is a bag rather than a set since the same unit can occur several times in a single sentence. It is partially ordered because a lexical unit may correspond to a discontinous string of tokens. Moreover, several lexical units may compete for the same string of tokens, e.g. there may be two competing analyses for a string such as *President Clinton*, one viewing it as the expression of a single unit and the other taking it as consisting of two units. For a single token to be associated with more than one lexical unit, however, the units must be regarded as alternatives. A single-word compound such as the previous example, *vattenreningsverk*, can not be analysed into its parts *vattenrening* (sewage disposal) and *verk* (plant).

The expressions of a lexical unit form a *paradigm*. A paradigm for a single-word unit includes its morphological variants. A paradigm for a multi-word unit should include syntactic variants as well. For instance, the lexical unit *turn_down* should include 'turned down', 'turning down' as well as expressions where the particle is separated from the verb by some appropriate phrase, as in the phrase *turned it down*. The current version of the system, though, only provides for morphological variants of regular single-word units.

As for the translation relation the basic assumption behind the LWA system is the same as for almost all approaches to word alignment, i.e. that *corresponding units of a translation lexicon have a greater tendency to co-occur in bitext segments than non-corresponding units*. Diffferent candidate translations for a given unit are rated by counting co-occurrences in segments and overall occurrences in the bitext as a whole and then an association score is calculated. The system currently supports three different association scores, mutual information, the Dice score and the T-score (Fung and Church, 1994). The latter is the one that has been used in all experiments reported in chapter 9 and is calculated as follows (with K as the number of bitext segments):

$$t = \frac{prob(V_s, V_t) - prob(V_s)prob(V_t)}{\sqrt{\frac{1}{K} prob(V_s, V_t)}}$$

The probabilities are estimated by means of the frequency counts. We have:

$$Prob(V_s, V_t) = \frac{occur(V_s \wedge V_t)}{N}$$

$$Prob(V_t) = \frac{occur(V_t)}{N}$$

$$Prob(V_s) = \frac{occur(V_s)}{N}$$

The probabilities indicate how likely it is to find V_s and V_t together in the same bitext segment, as well as finding V_s and V_t respectively in the text. N is here the number of bitext segment, containing a source side and a corresponding target side. The target candidate giving the highest score is selected as a translation provided the following two conditions are met: (a) the score is higher than a given threshold, and (b) the overall frequency of the pair is sufficiently high. These are the same conditions that were used by Fung and Church (ibid.).

It is further assumed that *a lexical unit in one half of a segment corresponds to at most one lexical unit in the other half.* This can be seen as a generalisation of the one-to-one assumption for word-to-word translation used by Melamed (1997a) and is exploited for the same purpose, i.e. to exclude large numbers of candidate alignments, when good initial alignments have been found. Thus, the system operates iteratively and after each iteration, tokens that have generated entries in the dictionary are removed from the bitext.

A rough partition of units and tokens into categories is employed to prevent spurious candidate pairs from being generated. A first division is made into *relevant* and *irrelevant units.* Irrelevant units are simply those that are not included in the set of candidates. They have to be listed explicitly. The reason for not including some items may vary with the purpose of alignment. Even if the aim is that the alignment should be as complete as possible, it might be useful to exclude certain units that are likely to confuse the algorithm. For instance, the *do*-support found in English usually has no counterpart in other languages. Thus, the different forms of 'do' may be excluded from consideration from the start. Apart from this restricted set of irrelevant units that can be known in advance, additions and deletions of lexical units in translation are taken to be haphazard.

Relevant units are further divided into *open class* and *closed class* units. Open class units can only be aligned with open class units, and closed class units can only be aligned with closed class units. Closed class units have to be listed explicitly, thus forming a sort of seed lexicon. Multi-word closed class units are listed separately. Closed class units are further classified for the purposes of alignment. The categories act as filters in the alignment process, excluding all candidate pairs that have non-corresponding categories. Standard parts of speech are employed, such as pronouns, prepositions and conjunctions for the purpose.

If some expression for the lexical unit U_T is found corresponding to some expression for the lexical unit U_S, then it is assumed that any morphological variant of U_T may correspond to U_S. This assumption is in accordance with the often made observation that morphological properties are not invariants in translation. It is used to make the algorithm more greedy by accepting infrequent alignments that are morphological variants of high-rating ones.

Some infrequent links are also captured by observing the occurrence of segments consisting of single tokens (*single-word segments*). If one half of a bitext segment consists of the expression of a single lexical unit, then assume that the other half is also. This is definitely a heuristic, but it has been shown to be very useful for technical texts involving English and Scandinavian, where terms are often found in lists or table cells (Tiedemann 1997).

A common assumption, which has been made by many researchers, is that *cognates* appearing on opposite sides of a bitext segment are translation equivalents (Simard et al., 1992; Melamed, 1995; Tiedemann 1997), see also section 2.5. These works have also demonstrated that cognate heuristics can be used with good effects on performance. Cognate heuristics are knowledge-lite in the LWA sense of the term and have also been implemented in LWA. One of the similarity functions used in LWA is based on the Longest Common Subsequence Ratio (LCSR) algorithm developed by Hunt and Szymanski (1977) and used for word alignment by Melamed (1995, 1999). The LCSR of two compared words is the ratio of the length of their longest common subsequence (LCS) divided by the length of the longest word.

$$LCSR(A,B) = \frac{length[LCS(A,B)]}{\max[length(A), length(B)]}$$

The cognate threshold when using LCSR as the cognate heuristic with LWA can be specified by the user (the default threshold used in the tests has been 0.6). Hunt and Szymansiki's LCSR algorithm will regard pairs like *scanner-skanner* as cognates (LCSR=6/7=0.857) whereas just comparing the first letters of the source and target word (Simard et al. 1992) would fail to recognise this pair as being cognates. Both types of cognate heuristics can be used in LWA.

Assumptions relating to *word order* are also commonly made. While word order is not an invariant of translation it is not random either. For this reason it is assumed that the candidate translations of a source unit can be found in roughly the same relative position as the unit itself. This assumption is implemented by two parameters: a link window and a position weight sequence. The user may determine the length of the link window and the position weights used for different positions of the window. Expressions that are close in relative position should receive higher position weights, while expressions that are far apart should receive lower weights.

7.1.2 Basic operation

The system takes input in the form of a bitext divided into segments. The current version requires the bitext segments to be numbered and the same numbers to be used as references on both halves of the bitext.

The algorithm is iterative, repeating the same process of generating translation pairs from the bitext, and then reducing the bitext by removing the pairs that have been found before the next iteration starts (Melamed 1997a). The algorithm will stop when no more pairs can be generated, or when a given number of iterations have been completed.

In each iteration, the following operations are performed:

(i) For each open class expression in the source half of the bitext (with frequency higher than the set value), the open class expressions in corresponding sentences of the other half are ranked according to their likelihood as translations of the given source expression. If the weight module is used, the position weights will affect the scores and the ranking.

This operation yields a list of translation pairs involving open class expressions.

(ii) The same as in (i) but this time with the closed class expressions. A difference from the previous stage is that only target candidates of the proper categories for the source expression are considered.

(iii) Open class expressions that constitute a sentence on their own (not counting irrelevant word tokens) generate translation pairs with the open class expressions of the corresponding sentence.

(iv) When all (relevant) source expressions have been tried in this manner, a number of translation pairs have been obtained that are entered in the output table and then removed from the bitext. This will cause fewer candidate pairs to be considered in the sequel and affect scores by reducing marginal frequencies and changing the contents of link windows. The reduced bitext is input for the next iteration.

7.1.3 Variants

The basic algorithm is enhanced by a number of options that implement the assumptions described above. An overview of the system is given below in Figure 16. The core of the system contains the alignment kernel that uses the word association score machinery to execute the basic processes of word alignment. In addition there are four main modules that can be invoked to improve the performance of the system. Apart from the main modules there are a number of parameters that can be set to determine what options should be used for a particular execution of the program.

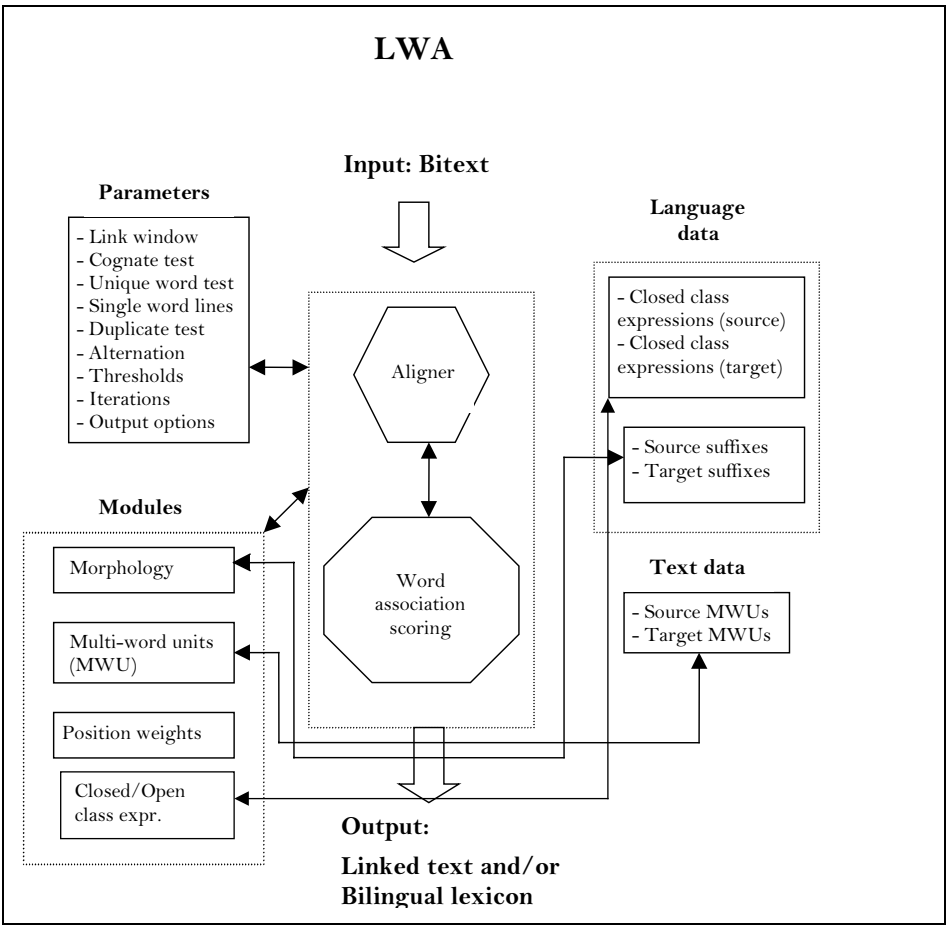


Figure 16. Overview of LWA

The four main modules are:

- A *morphological module* that groups expressions that are identical modulo the suffix sets;
- A *multi-word unit (MWU) module* that includes multi-word units generated in the pre-processing stage as candidate expressions for alignment.
- A *position weight module* that affects the likelihood of a candidate translation according to its position in the sentence; and
- A *closed/open class expression module* that makes the system observe the division into closed class and open class expressions and the subcategorisation of closed class expressions.

In addition to the above main modules, the following global parameters can be specified:

- *Link window.* The use of a link window will limit the search in the target segment. If a link window is used and its value is set to 5, this means that five words to the left and five words to the right of the starting position in the target segment will be tried as target candidates. For example, if there are 20 words in both the source and target sentences, and the source word under investigation is the 11th word of the source sentence, then a link window of size 5 will force the system to only consider the words from position 6 to 16 in the target sentence.

- *Cognate test.* If the cognate test is used, the user can choose between the Longest Common Subsequence Ratio (LCSR) test (Hunt and Szymanski (1977) and Melamed (1995)) or that the compared units start with *n* identical letters (cf. Simard et al. (1992)). The cognate test is then applied both in the Unique word test (see below) and as a heuristic when choosing among several candidates that show values above the word association threshold.

- *Unique word test.* When the unique word test is used, the bitext will be scanned for unlinked unique tokens (i.e., with frequency 1), and if there are unique candidates on both sides of a bitext segment, these words are then linked. The unique word test is used together with the cognate test, in order to increase the precision.

- *Duplicate test.* When the duplicate test is used, the bitext is scanned for duplicate sentence pairs, i.e., for recurrent sentence pairs where the source sentence and target sentence are identical. In translations with high recurrence degrees, this means that identical tokens of sentence pairs (source and target sentences) will be treated as a single instance.

- *Alternation.* If the alternation parameter is set to true, the linking process will be reversed at the end of each iteration, before the next iteration starts. In other words, when all possible links have been made from source to target, the system tries to find as many links as possible from the target to the source text. If alternation is used together with the morphology module, the possibility to link low frequency source expressions belonging to the same suffix paradigm is increased.

- *Frequency threshold.* This parameter specifies the lowest frequency used in the word association calculation.

- *Word association score threshold.* This parameter specifies the lowest threshold used for the word association score as well as what word association score to be used.

- *Number of iterations.* This parameter specifies the number of runs the linking process is executed.

The morphological module makes use of the simple suffix lists for each of the languages. The English suffix list looks as follows:

```
NOUN1:: {EMPTY, -s}                          # book, books
NOUN2a:: {-o, -oes}                          # tomato, tomatoes
NOUN2b:: {-ss, -sses}                        # mass, masses
NOUN2c:: {-sh, -shes}                        # brush, brushes
NOUN2d:: {-ch, -ches}                        # watch, watches
NOUN2e:: {-x, -xes}                          # box, boxes
NOUN3:: {-y, -ies}                           # lady, ladies
NOUN4a:: {-f, -ves}                          # wolf, wolves
NOUN4b:: {-fe, -ves}                         # wife, wives
ADJ1::{EMPTY, -er, -est}                     # short, shorter, shortest
ADJ2::{-y, -ier, -iest}                      # pretty, prettier, prettiest
VERB1:: {EMPTY, -ed, -ed} UU {-s, -ing}      # call, called, called, calls, calling
VERB2:: {-e, -ed, -ed} UU {-es, -ing}        # love, loved, loved, loves, loving
ADV1::{EMPTY, -ly}                           # love, lovely
ADV2::{EMPTY, -ally}                         # frantic, frantically
GEN1::(EMPTY, -'s)                           # Genitive singular. Frank's
GEN2::(-s, -s')                              # Genitive plural. teachers'
```

Figure 17. English suffix list used by LWA's morphology module

The effect is that words that start with the same character sequence are regarded as morphological variants if they fit *one* of the patterns described in the suffix list. All suffix patterns are tried on similar words of a candidate set (usually only the ten with the highest scores). If two or more variants are found on the target side, then these are treated as one set. For example, the VERB1 and VERB2 paradigms are suffix lists for regular English verbs. The morphological module collects open class translation pairs that are similar to the ones that are found by the basic algorithm. More precisely, if the pair (X, Y) has been generated as a translation pair in some iteration, other candidate pairs with X as the first element are searched. A pair (X, Z) is picked out if there exist strings W, F and G such that

$$Y = WF,$$

$$Z = WG$$

and F and G have been defined as different suffixes of the same paradigm. If several items Z are found, with suffixes from different paradigms, a selection is made, using the paradigm with the largest number of hits.

The *position weight module* distribute weights over the target expressions depending on their position relative to the given source expression. The weights must be provided by the user in the form of lists of numerical values (greater than or equal to 0). The weighted score for a pair is calculated as the sum of the position weights for the instances of that pair. This score is then used to adjust the co-occurrence statistics by using the weighted score instead of the co-occurrence score as input to the t-score formula. The threshold used is adjusted accordingly.

When the *multi-word unit module* is invoked, multi-word units are also considered as potential elements of translation pairs. The multi-word units to be

considered are generated in a special pre-processing phase and stored in a phrase table. The tool used here is the Frasse-1 or Frasse-2 system described in section 5.1.2. Along with the text-specific MWUs generated by Frasse-1/2, a number of general collocations contained in a language-specific list are searched for in the text. All general collocations found in the text are consequently merged with the text-specific MWUs found by Frasse-1/2 and used during the word alignment process.

T-scores for candidate translation pairs involving multi-word expressions are calculated in the same way as for single words. When position weights are used the position weight of a multi-word expression is considered equal to that of its first word.

It can happen that the t-scores for two pairs $<s,t^i>$ and $<s,t^e>$, where t^i is a multi-word expression and t^e is a word that is part of t^i, will be identical or almost identical. In this case the almost identical target multi-word expression is preferred over a single word candidate if it has a t-value over the threshold and is one of the top six target candidates. When a multi-word expression is found to be an element of a translation pair, the expressions that overlap with it, whether multi-word or single-word expressions, are removed from the current agenda and not considered until the next iteration.

If the *cognate test* is used, the list of open class translation pairs may be re-ranked. The highest-ranked target candidates for a given source word are inspected in order of decreasing scores and if one of them satisfies the cognate function, it is moved to the top.

The output from LWA comes in two versions: (1) the full-text alignment (see Figure 18), and (2) a bilingual dictionary (see Figure 19).

SOURCE: this CHAPTER TELLS YOU HOW TO SET UP MICROSOFT ACCESS ON A STAND-ALONE COMPUTER OR ON A NETWORK WORKSTATION
TARGET:I detta KAPITEL BESKRIVS HUR DU INSTALLERAR MICROSOFT ACCESS PÅ EN FRISTÅENDE DATOR ELLER PÅ EN DATOR i ett NÄTVERK

chapter => kapitel	(2 => 3)
tells => beskrivs	(3 => 4)
you => du	(4 => 6)
how => hur	(5 => 5)
to => i	(6 => 1)
set up => installerar	(8 => 7)
microsoft access => microsoft access	(10 => 8)
on => på	(11 => 10)
a => en	(12 => 11)
stand-alone => fristående	(13 => 12)
computer => dator	(14 => 13)
or => eller	(15 => 14)
on => på	(16 => 15)
a => en	(17 => 16)
network => nätverk	(18 => 20)
workstation => dator	(19 => 17)

Figure 18. Output from LWA: full-text alignment (linked tokens are shown in upper-case letters)

Source item	**Target item(s)**
/…/	
foreign minister	'utrikesministern', 'utrikesminister',
foreign policy	'utrikespolitik',
foreigners	'utlänningar',
foreman	'ordförande',
foremost	'främsta', 'främst',
forestall	'fransmännens',
forget	'glömma', 'glömmer',
form	'form', 'formen', 'utgör', 'gestalt',
formal	'formell',
formalities	'formaliteterna',
forms	'former', 'formerna', 'främst',
formulations	'formuleringar',
forth	'fram',
forties	'fyrtiotalet',
forty	'fyrtio',
forward	'fram',
fought	'bekämpat', 'bekämpade', 'slogs',
found	'fann',
founded	'grundades',
four	'fyra',
four thousand	'fyratusen',
/…/	

Figure 19. Output from LWA: bilingual lexicon

The full-text alignment data also come in a more system-oriented version where links are made as pointers to character positions.

The implementation was done in Perl with versions for Windows and Sun Solaris.

7.2 Adapting LWA to FRENCH/ENGLISH

A salient problem in the field of parallel text processing so far has been the lack of a common infrastructure for evaluation (see also chapter 8). For this reason LWA was entered in the word alignment track of the 1998 campaign of the ARCADE project (Véronis and Langlais 1999). Joining this multi-system evaluation would furthermore provide an opportunity to test the portability of the system, as the texts used in the ARCADE project were translations from French to English.

7.2.1 The ARCADE word alignment track

The task considered in the ARCADE word alignment track was that of *translation spotting*, a subtask of the full alignment problem. The objective in translation spotting is, given a specific source word or source expression, to identify the translation of all instances of this source word/expression in the target text. This is in a way reminiscent of information retrieval in that selected source words can be seen as queries to a set of documents. The corpus used for the experiment consisted of the French and English parts of the multi-lingual JOC corpus developed at Laboratoire Parole et Langage at Universite Aix-en-Provence. The French half contained some 1.3 million words and the English parts just over 1 million words. The corpus was delivered as two non-aligned text files. French words were selected on the basis of several criteria: (i) they should be polysemous in the bitext, i.e., they should have different translations in different contexts; (ii) they should be relatively frequent with an average frequency of around 60 in the corpus, and (iii) they should be adjectives, nouns or verbs in equal numbers. First 200 words were selected on the frequency criterion and these were submitted to a polysemy test by human judges. The twenty words of each category that came out as the most polysemous were selected for the experiment. Altogether they occur 3724 times in the corpus.

A group of human judges helped to create the alignments for the reference bitext. They were instructed to look for translational equivalents. Thus, if an occurrence of a word is not translated as a unit, but as part of a multi-word unit, this larger unit should be identified. Similarly, on the English side as many words as were needed to identify an equivalent should be marked. For instance, when the French word *claire* occurs as part of a comparison, the translation equivalent sought was *plus_clair – clearer* rather than just the pair of head words. Similarly, there are correspondences in the reference bitext such as *qui_comprends – including* and *conclure – lead_to_the_conclusion*. Multi-word units could also be discontinuous.

While the frequency range for the words should make them amenable to alignment, their polysemy and the need to take multi-word equivalents into account make the task difficult.

7.2.2 Adapting the LWA system to the ARCADE task

The ARCADE experiment posed several problems to LWA. The smallest problem was actually providing the system with the necessary French data, which was accomplished with the help of a French grammar in less than two days (13 man-hours). The English data files did not require any changes.

As the task of translation spotting is different from creating bilingual lexicons or terminology lists, it was decided not to extract monolingual multi-word units. The reason for this was mainly that the monolingual MWU extraction program (Frasse) would produce rigid phrases (proper names and terms that consist of adjacent units). In translation spotting (or full-text alignment) the task is to find whatever target unit that corresponds to a specific source unit, which means that the multi-word units in many cases are not of the type that Frasse could extract. For example, the French word **connaît** can correspond to a number of multi-word verb constructions in the corpus, such as **is familiar, is aware, will be aware** and **has been aware**:

> Source: La Commission **connaît** depuis longtemps…
>
> Target: The Commission **has** long **been aware**…

Therefore the best option was to exclude the multi-word processing stage and instead aim for linking as many of the individual words of a multi-word unit. In the example above **connaît** is linked to **aware**, which will result in a partial, overlapping link, instead of a strictly identical link.

In the training phase some bugs and shortcomings were also detected and corrected. As a result the system was made more general in several respects, allowing the user to control and vary system behaviour without changing the code. System changes and recoding took around 25 man-hours.

22 hours were spent on preparing the corpus. The corpus was aligned at (roughly) paragraph level with the aid of the DAVE toolbox (see section 5.3) and Perl scripts. The bitext segments were numbered as required by LWA. Moreover, a mapping had to be defined from ARCADE input format to LWA input format, so that the positions of every test word were known for both formats. A similar mapping had to be defined for the output formats. The corpus was then divided into a *kernel corpus* consisting of the segments where the test words were found, and a *peripheral corpus* consisting of all other segments. Apart from the actual work with the corpus and LWA, around 60 man-hours were spent on discussions and planning the ARCADE venture. In total, the conversion for LWA from an English/Swedish configuration to an English/French configuration took 115 man-hours. Due to the fact that the majority of this time was spent on configuring the corpus and meeting the

specific requirements of the ARCADE word track, conversion to other western European language pairs will be considerably shorter. The actual changes related to the new language required as mentioned above less than two days' work.

While the translation spotting task only required the system to find translations for the given test words, the interest from the Linköping team were more focused on the full-text alignment task. Thus, the aim was to align the whole corpus, but the workings of the system had to be changed slightly to fit the given task of translation spotting.

Two strategies were tested; each combining results from the kernel bitext and the peripheral bitext. The first strategy used the following steps:

1. Generate a translation lexicon from the peripheral bitext
2. Link the kernel bitext using the generated lexicon
3. Run LWA on the remainder of the kernel.

The second strategy was quite similar except for the fact that the two parts of the corpus were processed in a different order:

1. Run LWA on the kernel bitext
2. Generate lexicon from the peripheral bitext
3. Link remainder of kernel with the lexicon from previous step
4. Re-run LWA on the kernel.

An analysis of a sample of the output showed that the second strategy gave a higher recall but lower precision, therefore the first strategy was chosen.

In section 9.2, the results from the ARCADE word alignment track are presented.

7.3 Summary

In this chapter, the basic design and implementation of LWA have been presented. LWA is a knowledge-lite word-alignment system where statistical processing is combined with easily configurable linguistic resources. The portability of the system has been tested, illustrated by one system configuration for English/Swedish and one for French/English.

8 Evaluation of Word Alignment systems

In this chapter various approaches to the *evaluation* of word alignment are described. Evaluation is not only interesting when different systems are compared. For the system developers, evaluation is essential in order to know where and what to improve in a system. For users of word alignment systems, whether they are lexicographers, translators, terminologists or translation scholars, it is imperative to be aware of strengths and weaknesses of particular systems. The objective of this chapter is to look at the different ways a word alignment system can be evaluated. After arriving at some conclusions, the chapter concludes with a presentation of a piece of software, the PLUG Link Annotator, which can be used in word alignment evaluations.

8.1 Problems

There are several problematic issues for the evaluation of word alignment systems, the most important being,

- **The purpose of the alignment system**. A program designed for bilingual lexicon extraction differs from a program that aims at aligning a whole text with its translations on the word and phrase levels. Furthermore if the output data is used for bilingual concordance browsing, the system should be evaluated with this aim in mind.

- **Units**. What characterises the translation units? Should multi-word units be counted as such? Should function words be included or excluded?

- **Resources used**. When systems are compared, information on how long it takes to run the system on a particular bitext should be included, as well as extra resources such as bilingual lexicons and monolingual collocation lists.

- **Prior or posterior reference**. When alignment output is evaluated it can be compared to a Gold standard (sometimes referred to as *reference*), which is constructed *before* the actual alignment (*prior reference*), or experts can evaluate a sample of the output *after* the alignment (*posterior reference*).

- **Metrics and scoring method**. What metrics should be used? When the output is evaluated, there are several questions on how to judge partial alignments, deletions, insertions, segmentation errors and paraphrases.

- **Error analysis**. What is the nature of the mistakes that a particular system makes? Does it typically fail on certain types of collocations, on units within a particular frequency range, etc?

In the rest of this chapter, these issues will be addressed in relation to word alignment systems in general, but also to full-text alignment systems and lexicon extraction systems specifically.

8.1.1 The purpose of the alignment system

Sometimes it is difficult to distinguish between different types of systems which all share the general objective of identifying correspondences between text units in a source and a target text. However, a program that extracts a bilingual lexicon is primarily aimed at finding translations for content units, that is, terms, phrases and content words. On the other hand, one can say that a program that aims at aligning all tokens in a text can also produce a bilingual lexicon. The resulting bilingual lexicon (which is just a generalisation of all the token links) will typically contain entries that are not aimed for in a pure lexicon extraction program. The evaluation method should therefore be tailored to a specific type of alignment system in order to avoid unfair comparisons.

8.1.2 Units

In a pure word-to-word model (cf. Melamed 1995), many valid lexical units are missed due to the fact that they belong to collocations or complex paraphrases. For all kinds of word alignment linking, it is necessary to be able to handle multi-word segments in both the source and target text. Some approaches use pre-processing on only the source side (Melamed 1997b, Smadja et al. 1997) and then the target correspondences are estimated during the linking stage. In other approaches, both the source and target texts are pre-processed independently and candidate lists for both source and target multi-word units are created to be used in the linking process (cf. chapter 7 in this thesis).

The major difficulty is to identify all multi-word units present in a text, especially when the frequency is low. Furthermore, it is not obvious how to make the segmentation for certain multi-word units, such as particle verbs and idioms with variables (for example, "pull his/her/your/someone's leg").

Recall is also difficult to measure when multi-word units are considered, due to the fact that it is more or less impossible to know how many multi word units there are in a text. Recall measures can therefore in practice only be made on samples of a bitext.

8.1.3 Resources used

Some word alignment systems make use of extra resources, such as bilingual dictionaries, function word lists, morphological components, taggers, phrase lists or different separate programs for processing multi-word units. The resources used by a particular system are valid information in the evaluation. Even if a "black box" approach is adopted, and the output is judged against checked

reference data, the types of resources a system can utilise have to be accounted for if a complete picture of the system's performance is to be painted. Information on how long it takes to run the system on a particular bitext is also relevant for the evaluation as well as the platform and hardware that are used.

8.1.4 Prior or posterior reference?

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

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

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

Setting up a gold standard before the system is used, is definitely more efficient due to the savings in time. One gold standard can be used to check hundreds of sets of output data from one or several systems automatically. The drawback is that annotation guidelines as well as software for the annotation of the gold standard and the scoring have to be created, but once this is done, the advantages will outweigh the disadvantages.

8.1.5 Metrics and scoring methods

The standard metrics used for measuring the performance of NLP retrieval systems are *recall* and *precision*.

A standard way of describing recall would be "how often does the system provide an anwer?". Precison can be formulated as "how often are the answers correct?". If an alignment system produces alignments for 250 of the 500 entries in a gold standard, recall would be 50 per cent. The 250 found links can then be compared to the reference data and if 200 of these links were found to be identical to the reference, precision would be 80 per cent (200/250).

In the literature there are various approaches to describing and using precison and recall. Kitamura and Matsumoto (1996) regarded precision only on the

highest ranked *n* hundred candidates suggested by the system. Recall was then measured relative to the set of words that occurred at least twice in the corpus (i.e., the candidates above the frequency threshold built into the system). Gaussier (1998) used a similar approach where the top 500 links were checked for precision.

The standard recall and precision measures could be handled straightforwardly the links only consist of single words, but it becomes more difficult when the alignments are not one-to-one, which they indeed are not when multi word units are involved, as well as for deletions, insertions, segmentation errors and paraphrases.

The scoring for precision and recall can be adjusted to handle partial alignments by using some kind of proportional scores where each link will be judged depending on the number of words in system output and the reference data. In the ARCADE word alignment track (Véronis and Langlais 1999), precision and recall were tailored towards the translation spotting problem, that is the words in the reference were known beforehand. In the ARCADE project, the following measures for recall and precision were used:

Precision=Correctly proposed words/Proposed words

Recall=Correctly proposed words/Total reference words

In Table 53 below an illustration is given for three link instances. The first unit in the reference, "a/b/c" (i.e. a three-word phrase) is then compared with what the system proposed, namely "a/d". This gives a precision of 0.5 and a recall of 0.33 for this particular instance. If the system fails to propose a candidate when there exists an actual translation, both the precision and recall scoring will be zero. If a certain unit is not translated and the system also fails to find a link, the precision and recall figures will be 1 for such instances. The total precision and recall rates are then calculated as the average of all the link instances in the sample. In the ARCADE project, *F-measure* was also calculated. F-measure is the harmonic mean of recall and precision (see also section 8.8).

Table 53 Example of precision and recall scoring in ARCADE

Reference words	**Proposed words**	**Precision**	**Recall**
a/b/c	a/d	1/2=0.5	1/3=0.33
e/f	-	0	0
-	-	1	1
Average		0.5	0.44

The advantage of the above scoring method is that the successful linking of multi-word units is visible and rewarded, but also that partially correct linking is not deemed out entirely. One disadvantage of this approach is that the reference words and the proposed words are only considered from the point of view of the target. It presupposes that the source units in the reference are the same as the ones that the system has tried to link with corresponding target units. To

remedy this drawback, a comparison of the source segmentation between the reference and the alignment system could be performed.

However, the ARCADE precision score is also different from the standard precision score presented earlier, because the ARCADE precision is measured on *all* links in the reference data, including those where the system has not produced a result.

The ARCADE measures may suit the evaluation of translation spotting, but are not as useful for testing full text alignment. Translation spotting is a subproblem of full text alignment where the source expressions are known beforehand and the task is in principle constrained to finding the translations of these source expressions. For full-text alignment the reference data is seen as a representative sample of the full text, against which the system is compared for its capacity to align the full text.

8.2 Evaluation of full-text alignment

When the output of an alignment system is some kind of encoding of all lexical units and their corresponding translations, there are basically two ways to evaluate the output:

1. Comparison with a Gold standard that has been constructed in advance (prior reference).

2. Evaluation of a sample of the output after word alignment (posterior reference).

Gold standards for evaluating full-text alignment could come in different formats:

1. **Complete alignment of the sample.** This is a method where the source and target sentences in the sample are broken down into segments and the translation correspondences are marked. Melamed (1998a, 1998c) used this method in the Blinker project.
2. **The "spot check" method.** Here a number of words or expressions from the source text are chosen manually. All the sentence pairs that contain the singled-out tokens are presented to the annotator who chooses the corresponding target word/phrase. (Véronis 1998).
3. **Controlled sampling** of source words and expressions in the gold standard. This method is similar to the second alternative in that only a sample of source words are singled out. The difference lies in the way that the selection is made. With the controlled sampling method the selection is made by specifying how the sampling is to be done, for example, random sampling from the word tokens, sampling of only content words, sampling of word types instead of tokens, frequency-balanced sampling of tokens, etc.

The advantage of the first method is that nothing can be avoided. All the text segments in the sample have to be annotated. The disadvantage is that it can be hard to arrive at a single correct mark-up, especially when there are several annotators. A great deal of work therefore has to be put into creating unambiguous instructions that guide the annotators. Melamed's inter-rater agreement during the Blinker project was 75.92-85.11 per cent when comparing any two annotators and all annotations were compared. If function words were ignored the inter-rater agreement increased to 88.28-92.73 per cent (Melamed 1998c).

With the spot-checking and controlled sampling methods, it is possible to cover different types of words and phrases in a more consistent way. In the Arcade competition, 60 word types were singled out, 20 verbs, 20 adjectives and 20 nouns. Here all word types had a frequency of around sixty, but they were chosen on the basis that they exhibited some kind of interesting problem concerning polysemy. With the controlled sampling method could improve the evaluation of units over various dimensions, such as content vs. function words, frequency ranges, polysemy and parts of speech.

If evaluations of full-text alignment systems are to be really useful, it is not sufficient to know how well they perform in terms of precision and recall. An evaluation should also contain information on what the strengths and weaknesses of the particular system are. Therefore a predefined set of categories would help to describe the characteristics of the alignment

8.3 Evaluation of bilingual lexicon extraction

If the purpose is to evaluate an extracted lexicon, i.e. a set of link types, there are at least three ways to do this. For example:

- Compare the type links to an existing bilingual lexicon (the existing lexicon will function as a prior reference).
- Have the extracted lexicon evaluated by lexicographical experts (posterior reference).
- Measure the "explanatory power" of the extracted lexicon by applying it automatically on an already defined sample of the corpus (Melamed 1995, 1997a)

In some work presented in the literature, the explicit goal has been to extract bilingual dictionaries (e.g. Fung and Wu 1994, Fung 1995a, Fung 1995b, Fung 1998, Fung and McKeown 1996, and Kaji and Aizone 1996).

Using prior reference for evaluating bilingual lexicons requires an automatic comparison with a machine-readable bilingual dictionary. However, such a comparison could give the wrong results. Non-standard translations, translations of collocations, technical terminology, etc. are often not found in standard dictionaries, which, as a consequence, will produce misguiding scoring measures.

When posterior reference is used, i.e. when humans evaluate the extracted lexicons, the evaluators have to annotate each entry in the dictionary as correct, partially correct or wrong. By using this method, precision can be calculated either on the complete dictionary or, more practically, only on a sample of the extracted dictionary. In the first evaluation of LWA presented in chapter 9, this method was applied on all the source entries starting with three letters of the alphabet, N, P and O. The disadvantage of this kind of posterior lexicon evaluation is not only that each evaluation has to be redone every time the system has been run. It is also impossible to calculate recall, as there is no way of knowing how many entries that could have been extracted when multi-word units are included. Recall can however be measured in relative terms. In the lexicon evaluation presented in the next chapter (section 9.1), different configurations of LWA produced varying dictionaries of varying sizes, which makes it possible to compare recall in relative terms. Kaji and Aizone (1996) adressed the same problem by calculating *pseudo-recall*, which is the ratio of the number of correct entries in the extracted lexicon to the number of word types in either the source or target text. Another drawback of posterior reference applied to lexicons is that it is difficult to judge non-standard correspondences when the word pair is presented in a lexicon without its context. A type link may appear correct in the lexicon, even though it is based on erroneous link instances in the text. On the other hand, a non-standard translation can be judged as incorrect even though it is based on "real" correspondences in the texts.

A relatively simple automatic method has been suggested and used by Melamed (1995, 1997b) where one part of the corpus is put aside for testing. When the alignment has been done on the remainder of the corpus, the generated lexicon is applied on the test corpus and figures can be given on the proportion of matches found in the bitext segments. The assumption is that this method will be able to measure the explanatory power of the generated lexicon without any human annotation.

An alternative evaluation method could be a more pragmatic and practical approach, similar to the solutions suggested by Dagan and Church (1994) and Fung and McKeown (1996). Both adopt a way to measure the increase in efficiency that can be observed when translators are using a particular machine-extracted dictionary. The translators who tested the dictionaries extracted by Fung and McKeown, for example, increased the number of correct term translations by 47 per cent.

As an extension in the same vein, i.e. a practical kind of an evaluation, one could imagine a scenario where professional lexicographers use automatically extracted dictionaries to update commercial bilingual dictionaries. The lexicographical database that contains the commercial dictionary would be compared with the extracted dictionary and suggestions of possible new entries for the database would be presented to the lexicographers who in turn can choose whether or not the new entry should be added. Comparing how many entries that are actually added by using such a technique with the "old" way of updating dictionaries would prove a valuable evaluation of automatically extracted dictionaries. A glimpse of what such an approach could result in is given in section 9.4 in the next chapter, where the extracted lexicon from the Bellow novel was compared to a commercial English-Swedish dictionary and a number of new candidates not listed in the commercial dictionary are presented (see page 158).

The same information about the strengths and weaknesses mentioned for full-text alignment systems applies to lexicon extraction systems. Problems that arise from, for example, segmentation and stemming could be included in this set of pre-defined categories for lexicon extraction systems.

8.4 Word alignment evaluation– conclusions

Using reference data (gold standards) to check the performance of word alignment systems has several advantages, especially if different systems are to be compared or different configurations of one single system are tested. The reference data will make it possible to check system output automatically and in a consistent manner. The drawbacks are that dedicated software for creating the reference data as well as for checking the system output against the reference. However, it is possible that posterior reference could complement the gold standards, especially for lexicon evaluations. In such cases, human posterior evaluation of the extracted lexicon can capture features that cannot be captured by an automatic comparison to an existing lexicon.

Every evaluation should be accompanied with information on:

- the primary aim of the system (lexicon extraction or full-text alignment)
- the scoring methods,
- type of units (single-word – multi word units)

- extra resources (lexicons, taggers, multi-word unit extraction, etc.)
- information on the implementation and execution times.

Because of the advantage of of reference data, the PLUG[16] project decided to develop a dedicated software package for creating and evaluating word alignment output. The software package containing the PLUG Link Annotator and the Link Scorer is described in the next section.

8.5 The Plug Link Annotator

In this section the PLUG Link Annotator and the considerations that underlie its design will be presented. Primarily, it is developed with the aim of evaluating the word alignment programs used in the PLUG project, (Ahrenberg, Andersson and Merkel 1998) and Tiedemann (1998), but it is not tailor-made for these systems and could therefore also be used to evaluate other word alignment systems. In addition, the annotation guidelines are discussed. A scoring component, called the Link Scorer, which is used to produce automatic evaluations related to the PLUG Link Annotator, is also described

A gold standard is only one example of a product that can be generated interactively from a parallel corpus. For instance, if you are interested in studying the behaviour of a set of related words in translation, it would be useful to be able to make annotations in their concordances and generate a table or report of these annotations. The architecture of the PLUG Link Annotator facilitates extensions of this kind. Figure 20 below illustrates how different criteria can be used for selecting the setup of reference data.

[16] PLUG stands for Parallel Corpora in Linköping, Uppsala and Göteborg, a project jointly funded by NUTEK and HSFR under the Swedish National research program in Language Technology.

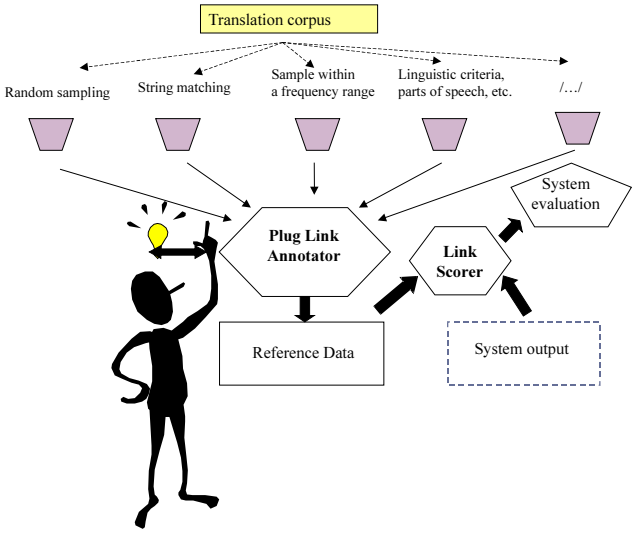


Figure 20. Overview of how the PLUG Link Annotator and the Link Scorer are used for evaluation of word alignment systems.

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

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

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

8.6 Using the PLUG Link Annotator

The PLUG Link Annotator is run interactively to create reference links, which can be used to measure the performance of a word alignment program automatically. The input to the PLUG Link Annotator is a list of source words together with the source sentences where they occurred and the corresponding target sentences. In the current version, experiments have been made with three types of selections: (1) a random selection of 500 words (as tokens) from each target text, (2) a frequency-balanced selection divided into five groups of 100 words each (frequency 1-2, 3-4, 5-9, 10-40 and >41) and (3) the same as (2) but excluding closed class expressions (that is, only content words are considered). The choice of input words could as mentioned above of course be made differently in the pre-processing stage.

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

The PLUG Link Annotator is provided with a web interface, with the underlying code as a Java applet. The Plug Link Annotator interface is shown in Figure 21 below.

The interface consists of four major fields:

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

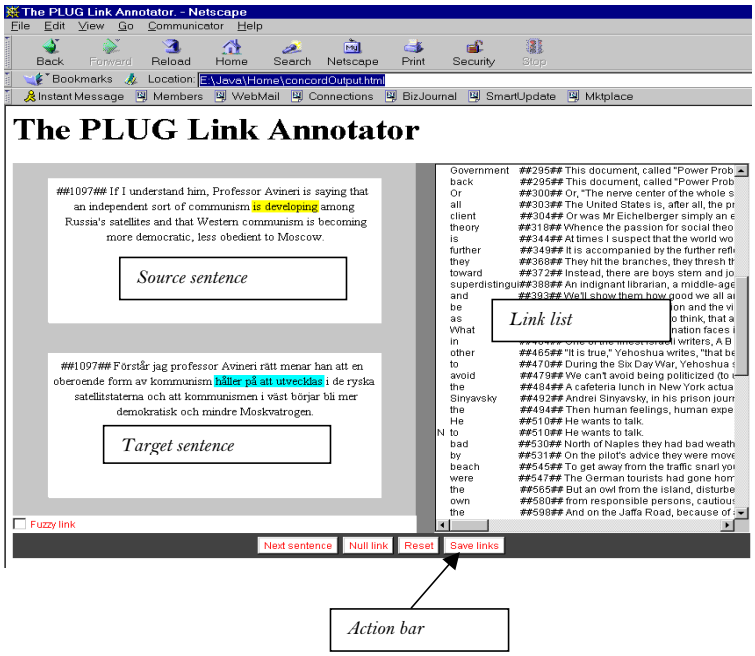


Figure 21. The Plug Link Annotator interface

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

8.7 Guidelines for the annotators

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

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

To ensure that there is a two-way equivalence, as many words as necessary

should be selected. Even if the starting point is always the source word, the selected parts in the two texts should correspond in both ways.

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

SOURCE: For more information on configuring a particular **SQL database server**, search Help for "ODBC drivers"...
TARGET: Mer information om hur du konfigurerar en viss **SQL-databasserver** finns i Hjälp under "ODBC-drivrutiner"...

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

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

8.8 The use of annotations and the Link Scorer

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

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

Figure 22. Entry in the output from PLUG Link Annotator

When a word alignment system's output is checked against the gold standard (the PLUG Link Annotator file), precision and recall figures are calculated automatically. The dedicated program for doing the scoring is called the *Link Scorer*. By scoring the results in this manner, it is possible to compare the performances of different systems. With data from the scoring phase, it is possible to pinpoint both strong and weak points of the systems, for example how the systems perform on multi-word units and fuzzy links.

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

An example of the output from the Link Scorer is shown in Table 54.

Table 54. Output from the Link Scorer

Number golden:	500
	(Regular: 388, Fuzzy: 26, Null: 86)
Number identical:	272 (r: 207, f: 2, n: 63)
Number partially linked:	109 (r: 100, f: 9, n:0)
Number completely different:	61 (r: 29, f: 9, n: 23)
Total number tried:	442
Number not tried:	58 (r: 52, f: 6, n:)
Recall:	0.884
Precision incl. partial matches:	0.862
Precision half partial:	0.739
F-measure:	0.805

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

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

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

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

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

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

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

8.9 Summary

There is considerable interest in text alignment on the word and phrase level, but some confusion on how alignment systems should be evaluated. The first, and perhaps most important, step is to decide the purpose and usage of such a system. If it is to be adopted for creating full-text alignments used for bilingual searches (bilingual concordancing) or for creating bilingual dictionaries, the evaluation must be tailored towards that particular usage. Secondly, the appropriate segmentation of the source text, in particular, is fundamental for comparisons of scorings between different systems. The metrics used for evaluating systems often vary between different approaches, even for systems with the same overall goal. The best solution, at least for full-text alignment systems, is the use of gold standards, where a correct reference is set up and against which the system output is measured. An interactive tool for creating reference data (the PLUG Link Annotator) and a program for scoring system output against the reference (the Link Scorer) have been presented. Further information, in addition to the scoring results, is also of interest and should be included in evaluations. This includes information on the type of errors that the system makes and also information about system performance in terms of time and memory usage as well as data on the implementation and hardware.

9 Evaluation of LWA

In this chapter three different evaluations of the LWA system are presented. At the end of the chapter a more pragmatic type of evaluation is presented and exemplified by comparing the LWA output to an existing standard bilingual dictionary. The first evaluation is the one reported in Ahrenberg, Andersson and Merkel (1998) when LWA was applied to two of the translations in the Linköping translation corpus. The first evaluation method focused on evaluation of the extracted bilingual lexicon. The second evaluation is the one done in conjunction with the ARCADE project described in the previous chapters. Here the ARCADE scores were used. The third evaluation reported in this chapter is performed on the same texts as the first evaluation, with the difference being that LWA had undergone some minor changes and that the evaluation technique took advantage of the PLUG Link Annotator (for creating a gold standard) and the Link Scorer (for automatically comparing the gold standard to the system output).

9.1 Evaluation 1 (English-Swedish using dictionary evaluation)

The first test of the LWA system was performed on two different texts as reported in Ahrenberg, Andersson and Merkel (1998). Here the basic findings of the experiment are summarised. Some basic facts for the texts are given in Table 55, while Table 56 presents figures for precision and recall. It should be noted that not all the LWA parameters and functions described in chapter 7 were implemented in this setup (for example, the LCSR algorithm in the cognate test and the unique word test).

The texts used were the novel by Bellow and the Microsoft Access User's Guide described in chapter 4. In Table 55 data on the source texts for the two bitexts are presented.

Table 55. Characteristics for the two source texts (the Bellow novel and the Access User's Guide)

	Bellow	**Access**
Size in running words	66,693	169,779
No of word types	9,917	3,828
Word types frequency 3 or higher	2,870	2,274
Word types frequency 2 or 1	7,047	1,554
Multi-word expression types (found in pre-processing)	243	981

It can be seen that the novel contains a high number of low frequency words whereas the program manual contains a higher proportion of words that the algorithm actually tested as the frequency threshold was set to 3.

Multi-word candidates were generated by the Frasse-1 system described in chapter 5. The table above illustrates that Frasse-1 identified around four times as many multi word units in the Access source text as in the Bellow text. The tests were run on a Sun UltraSparc1 Workstation with 320 MB RAM and took 55 minutes for the Bellow novel and four and a half hour for the Access program manual.

The evaluation method used was with a posterior reference, that is, the extracted bilingual lexicons were evaluated by a human judge. The task for the evaluator was to determine whether a lexicon entry (pair of source type and target type) were *correct*, *partially correct* or *wrong*. The evaluated sample from each lexicon contained source words starting with the letters *N*, *O* and *P*. Results from three LWA configurations are shown in Table 56 below. BASE indicates that LWA only used the statistical machinery and no other modules, AM-W that LWA used all modules except the position weight module, and AM that all modules were utilised.

Table 56. Results from two bitexts, using T-score only (BASE), all modules except the weights (AM-W), and all modules (AM)

	Bellow			**Access**		
	BASE	**AM-W**	**AM**	**BASE**	**AM-W**	**AM**
Linked source expressiones	1,575	2,467	2,895	1,631	2,748	2,878
Linked multi-word expr.	0	177	187	0	683	734
Link types in total	2,059	4,833	5,754	2,740	7,241	7,487
Links in evaluated sample	234	573	709	318	953	1,005
Correct links in sample	207	530	639	199	655	753
Errors in sample	21	19	30	51	137	122
Partial links in sample	6	24	40	68	161	130
Precision (incl. partial)	91.03%	96.68%	95.77%	83.96%	85.62%	87.86%
Precision (half partial)	89.74%	94.59%	92.94%	73.27%	77.18%	81.39%
Token recall	50.9%	54.6%	56.70%	60.2%	67.1%	67.3%
Type recall freq 3 or higher	54.88%	72.06%	82.65%	73.88%	82.10%	85.53%
Type recall freq 2 or 1	0	3.15%	4.87%	0	12.74%	12.74%

The results show that both recall and precision are improved considerably when the knowledge-lite linguistic data modules are used. The results for Bellow reveal that recall in relative terms is almost tripled in the sample, from 234 in the BASE configuration to 709 linked source expressions with the AM configuration. Precision values for the Bellow novel lie in the range from 91.03 to 96.68 per cent when partial links are judged as correct and slightly lower when partial links are scored as "50 per cent correct" (89.74-94.59 per cent). The use of weights seems to make precision somewhat lower for the Bellow novel, which perhaps could be explained by the fact that the novel is a much more varied text type.

For the Access manual, the relative recall results are as good as for the novel; three times as many linked source types for the AM configuration compared to baseline (BASE). Precision is increased, but perhaps not to the level that was anticipated at first. Multi-word expressions are linked with a relatively high recall (above 70%), but the precision of these links are not as high as for single words. A closer look at the links shows that one major problem lies in the quality of the multi-word expressions that are fed into the alignment program. As the program works iteratively and starts with the multi-word expressions, any errors at this stage will have consequences in later iterations.

Each module was also tested separately and this showed that each module would actually improve the results compared with the baseline configuration.

The results of this first simple experiment seemed encouraging and seemed to compare well with those reported in the literature, in spite of the knowledge-lite approach. This type of lexicon evaluation reflects some of the capacities of the alignment program on the general level. However, as the lexicon is a generalisation of the token links, it may be the case that one entry in the lexicon is a representation of only one token link whereas other entries stand for thousands of identical token links. Furthermore, every time a lexicon is produced it has to be evaluated, which makes this method very time-consuming. In the next two sections, LWA is tested with the aid of gold standards, first in the ARCADE way for French/English and then with the aid of the PLUG Link Annotator and the Link Scorer for English/Swedish.

9.2 Evaluation 2 (French-English the ARCADE way)

In the Arcade project the different contributions were evaluated with the aid of the reference text (see section 7.2.1 and 8.2) which was also distributed to all participants in the word track. With the aid of the reference it was possible for all teams to see where their programs had been successful and where there was room for improvement.

There were eight research groups that initially were interested in joining the word alignment track. In the end five systems completed the whole test, namely three systems from France (Xerox, CEA and LILLA), one from Canada (RALI) and the LWA system. The best performance came from the system from Xerox that uses taggers and bilingual dictionaries in its linking process (Hull 1998), i.e.

a knowledge-intensive system. The LWA ended up as third out of five teams when overall precision and recall were taken into account.

The five system results as they are described by Véronis and Langlais (1999) are presented below in Figure 23.

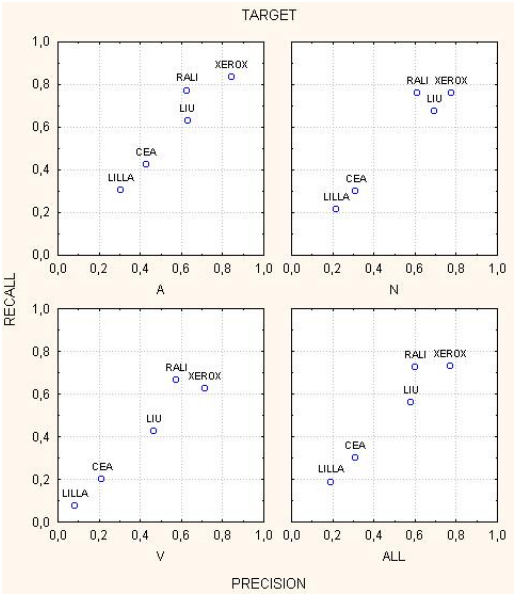


Figure 23. Arcade Word track results. LWA's results are indicated by "LIU"[18]

A comparison of the scores for LWA and the Xerox system is shown in Table 57 below.

Table 57. Arcade evaluation of the LWA system in comparison to the Xerox system

Category		Precision	Recall	F-measure
Adj	**LWA**	0.63	0.63	0.63
	XEROX	0.84	0.84	0.84
Noun	**LWA**	0.70	0.68	0.68
	XEROX	0.78	0.76	0.76
Verb	**LWA**	0.47	0.42	0.44
	XEROX	0.72	0.62	0.65
All	**LWA**	0.58	0.56	0.57
	XEROX	0.77	0.73	0.74

The above data show that the LWA performance is better for adjectives and nouns than for verbs, but it does not provide any answers to what is behind these figures.

In order to find out more about the strengths and weaknesses of the LWA system, the Arcade reference words and the LWA system output were used to

[18] The ARCADE evaluation is reported in Véronis and Langlais (1999) as well as on the web (http://www.lpl.univ-aix.fr/projects/arcade/2nd/word/results/charts.html)

extract examples of different types of links to see if more detailed information about the system's performance could be recovered.

First the reference words were compared with the output of LWA and all the *null links*, *overlapping links* and *strictly identical links* were identified. All links where the system failed were also extracted: the *no links* (where the reference specified a target and LWA had not identified a response) and *different links* (where the reference word(s) and the LWA suggestions were different). In Table 59 the number of links that LWA actually made and did not make are presented.

Table 59 Links made by LWA

Link type	Number
Null links	410
Overlapping links	132
Strictly identical links	1605
No links	1359
Different links	218
Total number of links	3724

The figures in Table 59 show the tendency for LWA to avoid making the explicitly wrong choices which is indicated by the relatively low proportion of *different links* and which yields a relatively high precision. On the other hand the system performs worse for recall because of the high *no link* figure.

By studying samples of erroneous links found among *overlapping links*, *no links* and *different links*, the following problematic cases were detected:

1. Multi-word units
2. Low-frequency words
3. Morphological variants
4. Overgeneration of good links
5. Size and segmentation of the aligned source and target sentences
6. Positioning of the link window.

A large proportion of the link mistakes stems from the cases where either the source or the target side contains a multi-word unit. As described in section 7.2.2, the MWU module was not used for the Arcade texts. Twenty-two per cent of the links in the sample contain multi-word units, which means that a full score for precision and recall was not attainable. Table 60 shows the number of links in the Arcade reference that contain multi-word units.

Table 60 Links with multi-word units

Link types	Number of links made by LWA	Incl. multi-word units (Arcade reference)	Proportion
Null link	410	1	0.00
No link	1359	551	0.41
Identical link	1605	69	0.04
Overlapping link	132	132	1.00
Different link	218	62	0.29
Total	3724	815	0.22

In the cases where LWA failed to identify a target word (*no links*), 41 per cent of the links contained a multi-word construction. When LWA proposed a different link compared to the reference, 29 per cent of these links had a phrase on either the source or the target side. If we exclude all links containing multi-word units, the precision score for LWA would be 90.78 per cent and the recall figure would be 58.16. In LWA there is no distinction between null links and no links, so the null links are not counted as links. If the null links were included, precision would increase to 92.57 per cent and recall to 72.22 per cent.

The second problem concerns the frequency of target candidates. Although the morphology module will assist to a certain degree in finding inflected variants with low frequency, there were many cases where the number of occurrences for a certain candidate did not reach the frequency threshold (which was set to 3). If the correct target word has a frequency of one or two or if the iterative linking already has established a number of links leaving a remainder of occurrences below the threshold, these links are never even considered. For example, it is impossible to find the pairs *courantes – mainstream*, *exceptionnel - one-off*, *fraîche – freshness*, and *utile – worthwhile*, because the English words all occur less than three times in the target text.

The third source of errors concerns the way that the system handles morphological variants. In the version of the system that was used for this corpus, the morphological variants are initially only invoked on the target side, which means that different inflectional forms of a target candidate are detected. When a source candidate is first tested, only the occurrences of that particular inflectional form are searched for. This means that there was an asymmetrical relation between the way morphological variants were handled. For example, the inflection **régulières** is not grouped together with its variants (régulier, réguliers, régulière). Instead the word form **régulières** is tried on its own, but does not reach a sufficiently high score to be linked. When the linking is reversed (from target to source), there is however a possibility that the different forms of the source word can be detected, but then some of the variants might already have been linked in previous iterations.

The fourth factor depends on the way the linking system is configured, the order in which words are tried and linked. In the setup used here, the corpus was divided into two sets; one *kernel corpus* and *one peripheral corpus.* The peripheral corpus was linked first and then a "dictionary" built from the peripheral corpus was used to link the core corpus before the main linking of the kernel corpus

started. This resulted in some overgeneralization in that some links made from the dictionary actually removed the actual target words that should have been linked in the main linking phase of the kernel corpus. Iterative processes have their advantages in that the search space is decreased as the process goes on, but if mistakes are made early on, there is no way to correct this at a later stage.

The fifth problem stems from the size and segmentation of the input to the word alignment program. The aligned bitext segments used as input were in most cases actually paragraphs, containing several sentences. The figures for both the peripheral corpus and the kernel corpus are summarised in Table 61. The fact that the peripheral corpus contains considerably shorter bitext segments on average is due to the fact to the large number of headings and section markers. In the kernel corpus there are very few examples of short headings.

Table 61 Length of bitext segments in the corpora

	Peripheral corpus		**Kernel corpus**	
	French	**English**	**French**	**English**
Average length:	25.54 words	20.87 words	71.38 words	57.31 words
Max. length:	315 words	264 words	369 words	269 words
Median length:	14 words	13 words	61 words	49 words
Total words:	1,058,296	864,598	227,763	182,868

If the aligned input text could be split up into shorter segments, the actual linking would have been both more accurate and more manageable. Many sentences have no natural correspondences and provide unnecessary noise for the word alignment program.

The final major source of errors has to do with the positioning of the *link window*. A link window of size 6 was used for both the corpora. From the sentence lengths on the target and source side respectively, it is clear that the English text is considerably shorter than the French original. The number of null links also indicates that information has been either deleted or compressed in the translation. This causes a problem for sentence pairs with large length differences.

Table 62 Positioning of the link window (source word = "exceptionnelle", target word = "outstanding")

L'octroi du mérite aux travailleurs est prévu dans la législation de plusieurs États membres. Sont ainsi récompensés certains travailleurs pour l'ancienneté de leurs services ou la qualité **exceptionnelle** de leur travail.	Awards for long services seniority and **OUTSTANDING** service are provided for in the legislation of **several** Member States.

Table 62 illustrates the problem of positioning the link window. The relative centre for the window is the word "several" and the correct target word

"outstanding" is placed outside the window which means that it is not considered as a candidate.

To sum up, LWA was successfully ported to another language pair with relatively small effort. The knowledge-lite approach does work in this respect, and bearing in mind that all the other systems came from French-speaking universities or companies, LWA stood up well in the competition. Three of the shortcomings listed above, namely the handling of multi-word units, linking of low-frequency words and the relatively large bitext segments, were addressed before the next major test of LWA was initiated. The third test is described in the next section.

9.3 Evaluation 3 (English-Swedish using a gold standard)

As a part of the PLUG project, it was decided that the translations in the PLUG corpora where some of the texts in the Linköping translation corpus also belong should be evaluated with the gold standards using the PLUG Link Annotator described in section 8.6 as a more objective way of comparing the system performance. In this section, two of the texts in the Linköping translation corpus are evaluated, namely *Microsoft Access User's Guide* and Saul Bellow's *To Jerusalem and Back*, i.e. the same texts as in the first evaluation. The first stage in the evaluation contained the following steps:

1. Create Gold Standards for the different translations in the Linköping translation corpus[19] with the aid of the PLUG Link Annotator.
2. Run the LWA system with a number of configurations of different modules.
3. Evaluate the different configurations with the Link Scorer.
4. Based on step 3, use the best module configuration and test what the effects are of changes to numerical parameters, such as frequency thresholds, initial frequency, t-values, position weights and size of the link window.

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

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

[19] The whole novel by Bellow was linked in this test, but the Access translation was shortened to approximately the same size as the Bellow translation in order to have comparable bitext sizes.

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

To limit the errors from the ARCADE experience, especially for low-frequency units, the frequency threshold was lowered to 2. This in conjunction with the introduction of an improved cognate test (the LCSR method, see section 7.1.1 and 7.1.3) as well the unique word test, would hopefully help LWA to link words with lower frequency than before, without sacrificing accuracy too much.

All the configurations used the same global parameters (when applicable):

- Break frequency: 10 (i.e., the initial frequency that the linking starts with).
- Frequency threshold: 2
- T-value threshold: 1.65 without weights and 2.5 with position weights.
- Link window size: 6
- Number of iterations: 8
- Cognate test: LCSR=0.6

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

Recall scores presented in the remainder of this section are calculated as was described in conjunction with the Link Scorer in the previous chapter (section 8.8). The measures for precision of two types: (1) *precision incl. partial*, that is, partial links are considered correct, (2) *precision half partial* meaning that a partial link is considered as "half-correct" (contributes with 0.5 for the precision score), see also section 8.8.

When the Link Scorer produced the results from all the configurations, it was obvious that the scores seemed to be very similar. Consider the data for recall, precision and F-measure in Table 63:

Table 63. Recall, precision and F-measure for nine configurations of LWA (using 500 random text tokens).

Text		BASE	SS	WS	FS	ALT	SING	PS	ALL	ALL-NOT-WS
Access	**Recall**	0.816	0.804	0.808	0.862	0.896	0.826	0.828	**0.884**	0.872
	Prec. incl. partial	0,838	**0,898**	0,824	0,865	0,814	0,837	0,833	0,861	0,871
	Prec. half partial	0,726	**0,784**	0,710	0,744	0,691	0,726	0,719	0,738	0,751
	F-measure	0,768	0,794	0,756	0,799	0,780	0,773	0,770	0,804	**0,807**
Bellow	**Recall**	0.630	0.590	0.692	0.688	0.672	0.668	0.654	**0.744**	0.738
	Prec. incl. partial	0,920	**0,955**	0,901	0,892	0,898	0,912	0,913	0,916	0,928
	Prec. half partial	0,842	**0,877**	0,815	0,810	0,817	0,824	0,836	0,815	0,828
	F-measure	0,721	0,706	0,748	0,744	0,737	0,738	0,734	**0,778**	0,755

The data in the table above do not correspond completely with the previous test of the Access and Bellow translations (section 9.1). The assumption was that the highest scores would turn up in the two rightmost columns (ALL and ALL-NOT-WS). The recall figures support this assumption; the ALL configuration has the highest recall (0.884 for the Access text and 0.744 for the Bellow text). Instead SS (baseline plus subcategorised closed-class words) contain the highest precision for both translations. The best scores for F-measure are found in the ALL configuration (Bellow) and in the ALL-NOT-WS configuration (Access). The scores seemed slightly mysterious at first, but when looking closer at them and also evaluating the size of the bilingual lexicons produced, it was clear that the data in Table 63 may not represent the performance of the different system configurations accurately. For example, recall for Access using the alternation configuration (ALT) is 0.896 whereas the recall using the ALL configuration is 0.884, that is the ALT configuration and the ALL configuration are almost identical. If the number of generated type links (bilingual lexicon entries) for these configurations is compared, the difference is, however, clearer. The ALT configuration produces a bilingual lexicon with 2,845 entries whereas the ALL configuration creates 6,770 lexicon pairs. The same pattern appears for both texts, see Table 64 below.

Table 64. Size of extracted lexicons for each configuration

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

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

To investigate if the problem was connected to the random text tokens present in the gold standard, the selection of source words was redone in a different manner. This time a frequency-oriented approach was used where the sampled source items were divided into five groups of different frequency (f): (1) f=1-2, (2) f=3-4, (3) f=5-9, (4) f=10-40, and (5) f>40, where each group holds 100 source tokens, in total 500 samples. The assumption here was that the preference for more or less only picking out high-frequency words would be avoided and thereby better represent the performance of the different system configurations. In addition it would provide a handle for observing the capacity of LWA in different frequency ranges. Table 65 below summarises the scores for recall, precision and F-measure for the total 500 source tokens in the gold standard.

Table 65. Recall, precision and F-measure for nine configurations (using frequency-balanced text tokens).

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

As can be seen in the table above, all scores are lower than when the random sampling of text tokens was used (see Table 63), but this is expected since the frequency-balanced gold standard will contain a higher proportion of low frequency tokens which are harder to align. However, using all the modules (ALL) is definitely the best option according to this gold standard; the ALL configuration receives the highest recall and F-measure. Using the subcategorised function words (SS) will actually produce a higher precision, but the SS recall is considerably lower than the ALL recall, which will favour making the ALL configuration as the preferred choice.

By looking closer at the different frequency ranges of the sample words in the gold standard, it is possible to observe where the strength and weaknesses of different modules lie. In Table 66 below, recall, precision and F-measure for the five different frequency ranges are presented for the ALL, BASE and SS configurations.

Table 66. Recall, precision and F-measure for three different configurations (frequency-balanced)

Text		f=1-2	f=3-4	f=5-9	f=10-40	f>40
Access	**Recall**	**0.460**	**0.850**	**0.78**	**0.88**	**0.89**
(ALL)	**Prec. (incl. partial)**	0.826	**0.835**	**0.910**	0.818	0.820
	Prec. (half partial)	0.685	**0.753**	**0.782**	0.739	0.719
	F-measure	**0.673**	**0.762**	**0.826**	0.766	0.745
Access	**Recall**	0.220	0.600	0.58	0.840	0.840
(BASE)	**Prec. (incl. partial)**	0.818	0.800	0.793	**0.857**	0.845
	Prec. (half partial)	0.727	0.717	0.690	0.756	0.744
	F-measure	0.338	0.653	0,630	**0.796**	**0.789**
Access	**Recall**	0.24	0.58	0.600	0.820	0.750
(SS)	**Prec. (incl. partial)**	**0.875**	0.828	0.817	0.854	**0.907**
	Prec. (half partial)	**0.750**	0.741	0.717	**0.756**	**0.793**
	F-measure	0.364	0.651	0.653	0.787	0.771
Bellow	**Recall**	**0.480**	**0.580**	**0.710**	**0.800**	**0.880**
(ALL)	**Prec. (incl. partial)**	0.979	0.983	**0.986**	0.913	0.932
	Prec. (half partial)	**0.917**	0.836	**0.887**	0.794	0.847
	F-measure	**0.630**	**0.685**	**0.789**	**0.797**	**0.863**
Bellow	**Recall**	0.050	0.270	0.530	0.750	0.900
(BASE)	**Prec. (incl. partial)**	**1.000**	0.926	0.925	0.907	0.900
	Prec. (half partial)	0.600	0.815	0.811	0.807	0.822
	F-measure	0.092	0.406	0.641	0.777	0.859
Bellow	**Recall**	0.050	0.230	0.480	0.680	0.780
(SS)	**Prec. (incl. partial)**	**1.000**	**1.000**	0.979	**0.941**	**0.936**
	Prec. (half partial)	0.600	**0.913**	0.885	**0.846**	**0.859**
	F-measure	0.092	0.367	0.622	0.754	0.818

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

Step 3 of the evaluation steps given on page 150 consequently resulted in that the ALL module being selected for further testing. The fourth step then involved

changing different parametric values for the ALL configuration. These parameters were:

1. Using a high break frequency, i.e., the initial frequency that the linking starts with at each iteration. Here the break frequency used was 1000 instead of the default of 10. This will force LWA to start linking high frequency words before the low frequency ones. (Break-High).
2. Using a low break frequency. Here the break frequency used was 2. This will force LWA to start linking low frequency words first (Break-Low).
3. Using a large link window of size 10 (instead of the default size 6) (Large-Window)
4. Using decreasing t-values in each iteration, i.e., in each iteration the t-value will be lowered by 0.5, thereby relaxing the constraints in each iteration (Decr-T-value)
5. Using a high t-value threshold in all iterations. Instead of 2,5 as the default t-value threshold, 4.0 was used (High-T-value)
6. Using a low t-value threshold in all iterations. Here 1.75 was used as the t-value threshold (Low-T-value)
7. Using "flatter" position weights, i.e. decreasing the effects of the relative positions (Flat-Weights)
8. Using "peak" weights, i.e., increasing the effects of the relative positions (Peak-Weights)
9. Using a high frequency threshold. Here 6 was used as the minimum frequency instead of the default 2. (High-Min-Freq)
10. Using a high t-value threshold without the weights module. Here 8.0 was used instead of the default 1.65. (High-T-No-Weights)
11. Using a low t-value threshold without the weights module. Here 1.5 was used instead of the default 1.65. (Low-T-No-Weights)

The results from the tests were that there were significant differences only in the following configurations: High-T-No-Weights and Low-T-No-Weights (i.e., when the position weights module was inactivated and either a much higher or lower t-value was used) as well as High-Min-Freq (when a higher frequency threshold was used). For the rest of the configurations the differences were very small for both the Access and Bellow texts. In Table 67 below recall and precision data for Access are shown.

Table 67. Recall, precision and F-measure for Access using varying parameters

Configuration	Recall	Precision (incl.partial)	Precision (half partial)	F-measure
ALL	0.772	0.842	0.736	0.753
Break-High	0.762	0.844	0.737	0.749
Break-Low	0.762	0.818	0.724	0.743
Large-Window	**0.776**	0.836	0.727	0.747
Decr T-value	0.772	0.837	0.731	0.751
High T-value	0.772	0.837	0.730	0.750
Low T-value	0.77	0.842	0.734	0.751
Flat-Weights	0.762	0.824	0.725	0.743
Peak-Weights	0.774	0.849	0.742	**0.758**
High-Min-Freq	0.63	0.842	0.731	0.677
High-T-No-Weights	0.338	**0.859**	**0.761**	0.468
Low-T-No-Weights	0.742	0.836	0.738	0.740

The recall figures vary between 0.338 (for the "High-T-No-Weights" configuration) and 0.776 (for the "Large-Window" configuration). The best results come from the configuration where the effects of the weights module were increased, i.e. that relative positions were regarded as more important ("Peak-Weights"). However, as can be seen above, precision (half partial) is preserved at around 75 per cent, with only minor differences between the configurations.

Recall and precision for the Bellow translation is presented in Table 68 below.

Table 68. Recall, precision and F-measure for Bellow using varying parameters

Configuration	Recall	Precision (incl.partial)	Precision (half partial)	F-measure
ALL	0.690	0.958	0.856	0.764
Break-High	0.684	**0.961**	0.857	0.761
Break-Low	0.688	0.954	0.852	0.757
Large-Window	0.684	0.960	0.856	0.760
Decr T-value	0.690	0.958	0.856	0.764
High T-value	0.678	0.958	0.854	0.756
Low T-value	**0.692**	0.959	0.855	**0.765**
Flat-Weights	0.69	0.953	0.851	0.762
Peak-Weights	0.686	**0.961**	**0.859**	0.758
High-Min-Freq	0.506	0.956	0.824	0.627
High-T-No-Weights	0.19	0.907	0.753	0.303
Low-T-No-Weights	0.60	0.952	0.837	0.699

The results for the Bellow translation are in relative terms similar to the data from the Access translation presented earlier. Recall drops drastically when the t-value threshold is increased and the weights module is deactivated (High-T-No-Weights). The same decrease in recall as for the Access translation is also

visible for the High-Min-Freq configuration (i.e., when the frequency threshold is increased to 6 instead of 2). And as in for the Access translation, the other parametric variations do not differ significantly from the ALL configuration.

A third kind of gold standard was also developed against the configuration where all modules and the default parameters were used (ALL) was tested. This time only content words were selected as input words to the Link Annotator. As for the second type of gold standard, the selection of words was divided into the five different frequency ranges. As can be expected, the selection of content words made recall decrease and precision increase. Recall and precision for the ALL configuration when they were evaluated against the three gold standards are shown in Table 69 for (a) random text tokens, (b) frequency balanced words and (c) only content words):

Table 69. Recall and precision for the ALL configuration as evaluated by three different gold standards

	Access		Bellow	
Gold standard type	**Recall**	**Precision (half partial)**	**Recall**	**Precision**
A. Random text tokens	0.884	0.738	0.744	0.815
B. Frequency-balanced	0.772	0.736	0.690	0.856
C. Only content words + frequency balanced	0.742	0.768	0.640	0.871

Consequently, recall and precision will vary depending on the type of gold standard used. Note that the recall and precision data in Table 69 are taken from one execution of LWA. The links and lexicons produced are therefore the same; it is the different strategies for selecting the reference data that are different.

So what does this tell us? Basically, the ALL configuration using the default parameters presented earlier is comparable to the parametric variations 1-8 presented earlier as far as recall, precision and F-measure are concerned. When the position weight module is used, it seems that this machinery is so powerful that it overrides most of the other parameters. It is only when the weight module is deactivated that changes to thresholds, such as minimum t-value and frequency, have any effects on precision and recall. Unlike the two versions of the MWU extraction program (Frasse-1 and Frasse-2), there does not seem to be a distinguishable feature that can be used in order to decrease recall in order to increase precision.

This third evaluation has shown the application of using the PLUG Link Annotator and the Link Scorer to automatically compare different configurations of a word alignment system. The scores for recall and precision can differ significantly depending on what kind of selection strategy is used for the input words to the gold standard. Here it has been shown that a random word sampling of source text tokens will not show the different strengths and weaknesses inherent in different configurations as clearly as a frequency-balanced sampling of input words. The reason for this is that a random text token selection will favour the selection of high frequency words, which in turn

are easier to align with a less sophisticated machinery. When more low-frequency words were included the relative differences between different setups of LWA appeared more clearly. To make the characteristics of a word alignment system (or configuration) even clearer, one could design other types of selection criteria, for example, word type based selection or selection based on grammatical criteria.

The use of prior references in the second and third evaluation can be complemented by data from the extracted lexicons. Information on how many lexical entries that have been extracted will shed a different light on the recall scores from the automatic scoring. For example, the data given in Table 64 (size of extracted bilingual lexicons) provide the information that type recall has more than tripled in the ALL configuration compared to the BASE configuration.

The improvement of LWA can be observed when the first and third evaluations of the Bellow text are compared. It was mentioned earlier that additional strategies such as a better cognate test and the unique word test improved LWA's capacity to link low-frequency items. Due to these enhancements, it was possible to lower the frequency threshold to 2 in the third evaluation of LWA, which increased type recall considerably and maintained high scores for precision. Although it is impossible to accurately compare a posterior lexicon evaluation with an evaluation made a with gold standard, the number of lexicon entries for the Bellow novel in the first test were 5.754 for the ALL configuration (see Table 56) and in the third test the ALL configuration produced 8,639 entries (see Table 64). (The Access text cannot be compared in the same way as the third test was only made on roughly half of the Access text used in the first test.) The conclusion is therefore that LWA has improved considerably since 1998 when the first test was done, in particular the capacity to link low-frequency words and expressions.

In the next section, the output from LWA has been evaluated, or rather applied, in a slightly different fashion compared to the usual presentation of recall and precision data.

9.4 Comparing LWA output to a bilingual dictionary

Dagan and Church (1994, 1997) as well as Fung and McKeown (1997) both measured the relative applicability and usefulness of terminology extraction by reporting the increased efficiency in the work of terminologists. This has not been done in conjunction with LWA and the DAVE toolbox, but one way of finding out whether the output from LWA could be useful for lexicographers working with bilingual standard lexicons is to compare the output from LWA with a standard bilingual dictionary. In this small test, the electronic CD-ROM version of the most comprehensive and widespread English/Swedish dictionary in Sweden was used, namely Norstedts Stora Engelska Ordbok (Norstedts 1996). The test was performed by manually checking a sample of the dictionary output from LWA when run on the Bellow novel. Two hours were spent on checking potential word pairs in the extracted LWA dictionary against the Norstedt dictionary. Proper names were excluded from this test, but the test

made it possible to identify a number of potential additions to the Norstedt dictionary. Consider the examples listed in Table 70 below.

Table 70. Eighty-six examples of word pairs extracted by LWA from the Bellow novel not found in "Norstedts Stora Engelsk-svenska ordbok"

English	Swedish	English	Swedish
ally	beskyddare	leaf cover	lövtäcke
ancient	uråldrig	loom	väveri
annihiliation	utplåning	low-quality	lågklassig
austere	kärv	machine gun	automatgevär
be obliged	nödgas	many	åtskilliga
clever	listig	mounted	beriden
convey	förmedla	municipality	myndigheter
crowded	fullproppad	moustachioed	mustaschprydd
desk	disk	news cast	nyhetsutsändning
disconcertingly	förvillande	newspaperman	tidningsman
discretion	gottfinnande	notch	remsa
dismember	stympa	occasionally	ibland
dissident	oppositionell	performance	uppvisning
dung-heap	dynghög	peripheral	bisak
ensue	vidta	pillbox	pillerdosa
equitable	resonabel	pretension	påstående
even	rent av	principally	i princip
even	ens	probably	antagligen
evenhandedness	opartiskhet	proclaim	förkunna
evidently	av allt att döma	punctilious	prudentlig
exceedingly	utomordentligt	rancour	bitterhet
exceptional	märkvärdig	reason	motiv
exploitation	utsugning	scold	förebrå
faith	tilltro	self-mutilation	självstympning
favour	förorda	self-worship	självdyrkan
fragrant	väldoftande	sinfree	skuldfri
freighter	lastångare	single	gemensam
frightful	förfärande	skyline	stadssilhuett
full	fyllig	style-designer	stilbildare
good-naturedly	glättigt	suit	kavajkostym
grand	storstilat	super	härligt
gun	skjutvapen	supply	förse
half-baked	halvbakt	swagger	skrävlare
helpless	underutvecklad	terrible	förfärande
hence	alltså	terrible	gräslig
hence	sålunda	thrilling	tjusande
hillside	sluttning	trouble	osämja
holocaust	utrotning	unclassifiable	obeskrivbara
hunger	svält	unlovely	oskönt
illusory	en illusion	vaulted	valvtäckta
impose	tillämpa	wearing	iförd
insertion	inträde	vexation	vända
knowledgeable	initierad	virtuous	upphöjd

What is illustrated above is the absence of the *pair* of a particular English and Swedish word in the Norstedt dictionary. Most of the English words were found in the Norstedt dictionary, but with other translations than the ones presented above. Some of the examples are perhaps slightly outdated and would not be considered for inclusion in a future edition of the dictionary, such as *suit-kavajkostym*, *newspaperman-tidningsman* and *freighter-lastångare*. However, I was surprised to find that pairs like *probably-antagligen*, *terrible-gräslig* and *occasionally-ibland* are not found in the Norstedt dictionary. What this minor test shows is that it is not only peripheral and domain-specific words that can be extracted with a system like LWA, but also words that most people regard as being part of a general bilingual vocabulary. If this kind of application was used on large volumes of parallel texts, I think the possibilities for lexicographers to find a solid empirical foundation for what to add (or possibly remove) from existing dictionaries would improve drastically. Bilingual concordances could also be of help to lexicographers when they want to test a given hypothesis, for example, if a certain word has been used or translated in a certain way. However, with bilingual concordances the user has to know in advance what she is looking for, be it a source expression and/or a target expression. With word alignment systems, specific hypotheses are not required. The systems will find a set of significant results and present them to the user, who in turn can, either manually or with the aid of software, start to filter the results and find relations that had been missed before.

In the future we may see applications where parallel corpora are used together with word alignment systems to automatically compile dictionaries that are then compared automatically to existing dictionaries. Furthermore, the parallel corpora will provide a great resource for describing the actual use of lexical items, especially in translations, which could be used for illustrations of usage for lexical entries (cf. the use of corpora in the work with the COBUILD English dictionary (Sinclair 1987, 1995) and the development of a French Canadian/English dictionary (Langlois (1996)).

9.5 Improving the system

The multi word unit module included in LWA is primarily dictionary-oriented; i.e. it can detect and link rigid adjacent multi word units, particularly noun phrases. In the Arcade project, the task at hand was construction-oriented, i.e., there was a need to identify corresponding constructions such as *se poursuit* and *is continuing*. A construction-oriented handling of multi-word units calls for a knowledge-intensive approach, including tagging and lemmatisation of the texts. With the knowledge-lite approach present in LWA, the extraction of all possible constructions would result in an enormous number of construction candidates from the bitext, which would give rise to the generation of too many irrelevant units. An alternative way of handling multi word units could be to first run LWA without the phrase module and link as many single word units as possible. Then a new variant of the MWU module would be applied, taking source MWUs and the hitherto linked bitext as input. The MWU module would then use the single word links as clues in how to establish correspondences between

the source MWUs and different candidates in the target that are calculated in the manner suggested by Melamed (1997a). Such an approach would mean a departure from the path of iterativeness, as previous links may have to be broken up in order to find the most probable links for a certain MWU.

The handling of low-frequency units and morphological variants were partly addressed before the third evaluation by including the unique word test and improving the cognate test. Furthermore, the third evaluation showed that it was possible to lower the frequency threshold to 2 (instead of 3 which was used in the first evaluation) and thereby increase recall without sacrificing precision. Additional improvements could to some degree be made by a symmetrical handling of linking morphological variants. At present, LWA fails to group together morphological variants on the source side at the initial stage because the morphological patterns are not tested for the source unit. By improving the way the morphology module is invoked, the source and target variants could be tested at the same time and thereby improving the possibilities to correctly link source words with low frequencies. If this is to be done within a knowledge-lite framework, without using lexicons and morphological analyzers, then a more sophisticated statistical method for clustering variants belonging to the same inflectional pattern has to be developed.

The size and segmentation quality of the input to the word alignment program is crucial for the results. In the Arcade attempt, problems arose because of the length of the segments in the bitext. A link window that helps to decrease the search space was used, but the positioning of the window in the target segment may be wrong as the placement is based on the relative position of the source word. The positioning of the window could be improved by utilising already established, "safe" links from previous iterations (cf. Melamed, 1995).

A better traceability feature in LWA would make evaluations easier. In order to understand what happens during the linking process, each link could be annotated with a stamp that signals when and how the link was made. For example, it would be possible to accurately pinpoint the causes for link failures if there was information about the number of available target candidates at a certain time, whether a certain link was made as a cognate link or with the "single-word line" strategy.

9.6 Summary

In this chapter LWA has been evaluated using three different methods: (1) lexicon evaluation, (2) translation spotting (the Arcade way); and (3) using the PLUG Link Annotator. It can be concluded that using reference data (as in the second and third evaluation) has advantages in that many more types of evaluations can be made when different systems or different configurations of a system are to be compared. However, the choice of selecting what words that should go into the reference data will influence the scores for precision and recall. Therefore it is not possible to compare precision and recall scores when different types of reference data have been used. The third evaluation showed that it was possible to single out an optimal configuration of LWA by using

prior reference data. It also showed the relative strengths of the different modules in the system. The evaluations reveal the improvements in LWA between the first and the third evaluation. The improvements are especially visible when it comes to linking low-frequency units. The latest version of LWA is therefore considerably better as far as recall is concerned while precision is maintained at a high level.

LWA is a knowledge-lite system in the sense that it does not require knowledge-intensive resources, such as lexicons, parts-of-speech information, grammars, etc. The knowledge-lite approach has its advantages when it comes to changing language pairs or text domains. In knowledge-intensive approaches to word alignment, each new language configuration has to be provided with large amounts of linguistic information. LWA was first developed for English/Swedish, and later LWA was rapidly ported to French/English for the ARCADE word alignment track. LWA ended up in third place in the ARCADE evaluation behind two knowledge-intensive systems (both developed by French-speaking research groups). A knowledge-lite approach is not an end in itself, but for languages where linguistic resources are limited, or when a swift adaptation to a new domain is required, it may be the only solution.

10 Translation correspondence – a model

One of the major fields in translation studies is the study of corpus-based analysis of text type-specific source and target texts where translation correspondences are uncovered and described (Koller 1983). This field involves a comparative study of syntax, semantics and style in the corresponding texts. It also entails development of text type topologies and contrasting the different languages as far as style and text norms are concerned.

Looking at the quantitative data from the translation corpus described in earlier chapters, there are obvious limitations in what kind of conclusions can be drawn regarding the actual translations. A marked-up translation corpus with linguistic features would make a much more interesting corpus to explore, given the right tools. However, even a parts-of-speech tagged corpus would not contain information about structural shifts that take place in the translation process. Nor would it give any information on to what extent content is preserved or modified.

In chapter 2 (section 2.2.4) translation operations are discussed in a general sense. And as Toury (1995) has pointed out, it is necessary to identify relationships between pairs of source text and target text segments in order to be able to make any generalisations or draw any implications for the underlying concept or norms behind translation. It is the identification of such relationships that is the topic of this and the following chapter.

Translation operations or translation shifts have been the subject of other studies. Recently Munday (1998) published a study of a small set of translation shifts occurring in a translation of eighteen pages of prose translated from Spanish to English. In the late eighties, van Leuven-Zwart (1990) developed a model designed for the description of fiction translations. Among van Leuven-Zwart's conclusions when the model had been used to analyze literary works into Dutch from other European languages, were that the studied translations of fictions were more specific and that they contained more explanations than the originals. Furthermore she found that the translation strategy was more towards target-orientation than source-orientation. The tendencies that the translations were more specific and the relatively higher presence of explanations are in line with Baker's (1993, 1995) proposed universals of translation (cf. section 2.2.5).

In automatic translation systems, structural translation shifts are necessary if the generated translations are to be grammatically well-formed. In most systems only obligatory grammatical shifts can be handled at best, and more complex

shifts have to be ignored. The Logos MT system for English-German (LMT) seems however to include some capacities for taking care of more complex structural shifts, based on context and lexical data (cf. Gdaniec and Schmid 1995, Gdaniec 1998).

To be able to study translation shifts, a correspondence model was developed that included a set of translation-relevant linguistic attributes. The model takes its starting point from Wollin's and Platzack's work in the eighties (Wollin 1981, Platzack 1983, see also section 10.2.1.1 later in this chapter), but is slightly modified from their approach as it aims at describing changes both in structure and in content. The current model has then been applied to a sample of source-target sentence pairs from the Linköping Translation Corpus. The hypothesis was that a more in-depth analysis of the translations from the translation corpus would improve the possibilities of capturing the characteristics of the different translations. In this chapter a translation correspondence model is outlined and in the next chapter the model is applied to a sample of the translations in the translation corpus.[20]

10.1 Objectives

The objectives of the correspondence model and its application to the translation corpus were to find out more about the exact functional/structural changes occurring in the translations as well as any changes in the content expressed linguistically.

The questions in search of an answer were the following:

- Do the structure and content present in the source change in the translation? If so, what characterises these changes?
- Are there any differences between the different text types in the corpus as far as structural and semantic changes are concerned? If so, what characterises the translations of different text types?
- Are there any systematic patterns between structural patterns in the source and the target?
- Does the translation method (human translation, translation memory-based translation and automatic translation) influence the characteristics of the translations?

10.2 Method

In this study a number of measures have been developed and used that in different ways try to capture structural and content-related correspondences between source and target texts. To be able to use the measures, different objects and translation operations have to be defined. The measures were applied on

[20] Chapters 10 and 11 are a revised and extended version of Ahrenberg and Merkel (1997).

English source texts and their Swedish translations, but most measures can be applied to other language pairs, with minor modifications.

10.2.1 Syntactic and semantic correspondence

The focus in this section is to try to quantify structural and content-related relationships between source and target texts. Due to the time-consuming nature of such analyses, the study is concerned with samples taken from the different texts. The number of samples from each text in the study has been limited to 100 pairs of text sentences, all translated 1-to-1. The selection of these pairs has been made randomly.

A detailed system of operations to describe differences between source and target sentences has been presented by Wollin (1981). Platzack (1983) has used this system in a somewhat simplified form where he applies it to English source texts and Swedish target texts. In Larsson and Merkel (1994) the Wollin system was applied to a Finnish-Swedish translation. Hasselgård (1996) made a similar study of novels translated from English to Norwegian and Norwegian to English.

10.2.1.1 Platzack and Wollin

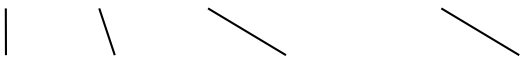
Platzack looked at six Swedish translations of Lewis Caroll's *Alice in Wonderland*. By classifying the syntactic correspondences between the English and Swedish texts as the result of eight basic operations (taken from Wollin 1981), he compared the distribution of these correspondences between the different translations. The operations were the following:

Structural identity, meaning that the syntactic structure is preserved in the translation, down to the primary constituent level.

1. Structural identity

The syntactic structure in the two sentences is identical, down to primary constituent level. Changes inside a primary constituent (agent, subject, finite verb, objects, etc.) are not recorded nor if there are changes in the order of constituents to the right of the finite verb

The long grass rustled at her feet as the White Rabbit hurried by.



Det höga gräset prasslade vid fötterna på henne när den vita kaninen rusade förbi.

2. Transponation

One primary constituent is before the finite verb and one is after the finite verb in the respective sentences.

The next moment she felt a violent blow underneath her chin

Strax kände hon ett hårt slag under hakan

3. Function adjustment

A primary constituent gets a different syntactic function in the target sentence.

and the Dormouse followed him

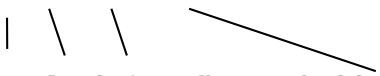
och hasselmusen följde efter

The direct object "him" is in the translation changed to a particle.

4. Addition

A constituent is added in the translation.

He moved on as he spoke.



Han flyttade fram till nästa stol vid dessa ord.

The adverbial "till nästa stol" has been added in the translation.

5. Deletion

A constituent is deleted in the translation. Below the English "to her" is not translated into Swedish.

and they repeated their argument to her

och var och en kom med sina skäl

6. Divergence

One source constituent is split into two constituents in the target sentence.

she could not remember ever having seen such a thing

hon kunde inte komma ihåg att hon nånsin hade sett det

7. Convergence

Two source constituents are combined into one constituent in the target sentence (the inverse of divergence).

He moved on as he spoke.

Han flyttade *medan han talade*

8. Combination of convergence and divergence

One constituent is the subject of both convergence and divergence

The Queen's argument was that ...

Drottningen påstod att ...

Between 259 and 275 operations were identified in each of the six translations. The majority of the operations belonged to the structural identity class (between 58 to 75 per cent). This number indicates in a syntactic sense how "free" a translation is from the original, although there may be other factors involved as well.

10.2.1.2 Hasselgård

Hasselgård (1996) made a text-based, contrastive investigation of word order in Norwegian and English, of translations in both directions. She analysed a small set of sample sentences in three translated novels using the typology of linguistic correspondences developed by Dyvik (1993)[21]:

> **Type 1**: Word-for-word correspondence, tolerating morphological discrepancies
>
> **Type 2**: Each content word in the SL sentence corresponds to a content word in the TL sentence, but with a different order between constituents
>
> **Type 3**: At least one source content word does not correspond to a target content word, but the logical form of the interrelated expressions is the same
>
> **Type 4**: None of the above – the translation is some kind of paraphrase.

Hasselgård reports the following distribution in the three novels:

[21] Dyvik used this typology to describe correspondences between Norwegian and Swedish.

Table 71. Types of word order correspondence in three corpus samples (Hasselgård 1996)

	No. of sentences	Type 1	Type 2	Type 3	Type 4
Novel 1 (Eng.-Nor)	26	12	8	6	-
Novel 2 (Nor.-Eng.)	26	11	5	10	-
Novel 3 (Eng.-Nor)	26	14	8	4	-

In her small study, 26 sentence pairs in each novel, Hasselgård has extended Type 1 to include sentences in which the clause constituents appear in the same order, but where the internal structure of the constituents may be different. According to Hasselgård, the use of the general types 2 and 3 causes problems in languages such as English and Norwegian if you want to "handle the contrasts in a way that is systematic and revealing". Types 2 and 3 should be broken down into subcategories (fronting, inversion, adverbial placement, etc.) where the type of difference and possibly also the cause could be identified. There were no translations of type 4 in the sampled sentences, but some examples of this type were present elsewhere in the texts.

10.2.2 Differences in this approach

Wollin's and Platzack's systems are absolute in the sense that they record all structural and functional differences between the source sentence and its translation, independently of whether they are based on grammatical differences between the languages or are optional choices on part of the translator. In this study Platzack's system has been taken as the starting point, but modified for the particular objectives of this investigation. The main difference is that the system described here is relative, i.e., structural changes that stem from necessary grammatical modifications are not recorded. The reason behind this is a wish to focus on how the choices made by the translator are influenced by the kind of translation tool she is using. The pure contrastive aspects between source and target language are therefore considered to be less in focus in this study. One assumption is that translators rarely breaks the grammatical and stylistic norms of the target language, especially when the translation concerns commercial technical texts translated by professional translators.

Drawing a borderline between linguistically necessary and optional changes is naturally not easy. The most important motivation in this study has been to arrive at a consistent classification and evaluation of the existing structural changes in the texts. The obligatory grammatical changes are therefore not as interesting as the changes that are voluntary on the part of the translator from a stylistic viewpoint.

Some clear cases of necessary changes between source and target that are not recorded in the Linköping model are the following:

- Differences in word order due to the fact that the finite verb in a Swedish main clause is positioned in the second position of the clause;

For Bray himself, it was an absence somehow always present

För Brays vidkommande var det en frånvaro som på något sätt hela tiden var närvarande.

- Functional changes due to valency differences between the main verbs, e.g., the use of a prepositional object instead of a direct;

...by selecting Dual Boot from the Command Prompts folder

...genom att dubbelklicka på Dubbla startsystem i mappen Kommandosessioner

- Genitival constructions where the English *of* is translated by *s*;

We are living in the age of science.

Vi lever i vetenskapens tidsålder.

Cases that are less clear-cut, but which are still considered to be necessary are for example the following (they are at least strongly (negatively) stylistically marked):

- The movement of an adverbial subordinate clause from the initial position to the position after the verb in a Swedish imperative sentence;

If you are not sure whether to partition your hard disk, select the default.

Välj standardvärdet om du är osäker på om du behöver partitionera hårddisken.

- The use of *inte* as a translation of the negative *no* in certain positions;

...he could find no French writers who would talk to him.

...kunde han inte hitta några franska författare som ville tala med honom.

Some frequent cases analysed as optional changes are the translations of an English imperative with a Swedish modal verb plus infinitival verb, and Swedish passive constructions which are used as translations of English active sentences with an instrumental subject:

If you are installing multimedia support from a network, access the network where the multimedia programs reside.

Om du installerar från LAN måste du logga in på det LAN där programvaran för multimedia finns.

Paradox stores important information about a table's primary key in an index (.PX) file. (MM)

I Paradox sparas viktig information om tabellens primärnyckel i en indexfil (.PX).

Th rule of thumb in the classification process has been the following: If a word-for-word, stylistically correct Swedish translation which is structurally closer to the source than the given translation is, cannot be found, then the given translation has the highest possible structural correspondence to the original. In the last of the two examples above, the literal translation *"Paradox sparar viktig information..."* would be considered structurally closer to the source than the actual translation. When the relation of "highest possible structural correspondence" holds between the source and target, this relation is called *isomorphic*.

The optional structural changes that can occur in the translations are analysed in the following categories and subcategories:

1. Changes related to the function and properties of clauses:
    - Voice shift (e.g., active > passive)
    - Sentence mood shift (e.g., imperative > declarative)
    - Shift of finiteness e.g., finitival > infinitival verb construction)
    - Level shift (e.g., main clause > subordinate clause, clause > phrase)
    - Function shift (e.g., temporal clause > conditional clause)
2. Changes related to the function and position of constituents:
    - Function shifts (e.g., manner adverbial > predicative)
    - Level shifts: (e.g., phrase > clause)
    - Transpositions (changes in order between constituents)
3. Changes related to the number of constituents:
    - Additions
    - Deletions
    - Divergences (one source constituent > two target constituents)
    - Convergences (two source constituents > one target constituent)
4. Paraphrases, which influence at least two constituents and cannot be split up into several smaller changes.

Changes of type (1)-(3) are regarded as *simple* while paraphrases are inherently *complex*. Note also that the simple changes can involve several constituents. A shift of sentence mood from imperative to declarative implies that a subject (implicit in the source) has been made explicit in the target. This leads to an addition of one constituent in the translation, but because this change can be seen as a necessary implication of the mood shift, this particular type of addition is not recorded separately.

Each sentence in the sample has been given an analysis in *translation segments*. A translation segment is defined as a nucleus, which is constituted by a content word, or a multi-constituent term, and its accompanying functional words.

Furthermore, main and subordinate clauses as well as a variety of phrases are also identified. Phrases are categorised based on their syntactic function in the clause. For each sentence pair the number of translation segments and different kinds of changes are recorded. The example above,

If you are installing multimedia support from a network, access the network where the multimedia programs reside.

Om du installerar från LAN måste du logga in på det LAN där programvaran för multimedia finns.

is given a structural description which among other data contains the following information:

Table 72. Structural description including number of translation segments and changes from source to target

Translation segments, source:	*10*
Translation segments, target:	*11*
Mood shift (Imperative > Declarative):	*1*
Addition (Verb):	*1*
Deletion (DirObject):	*1*

10.2.3 Measures

At this point the first measure for correspondence can be given. A simple measure that indicates the proportion of isomorphic translations is defined as:

$$\textit{Isomorphic index} = (I \times 100)/N$$

where I is the number of isomorphic sentence pairs and N is the number of sentence pairs in the sample. A slightly more fine-grained measure can be arrived at the if the pairs that are not isomorphic are divided into different categories. A translation can be called *semi-isomorphic* if it either (i) contains one simple change at the most, or (ii) contains at least seven translation segments and involves two simple changes at the most.

A translation is considered to be *heteromorphic* if it contains either (i) one paraphrase, or (ii) at least three simple changes, or (iii) exactly two simple changes and less than eight translation segments in the source sentence(s).

The borderline between semi-isomorphic and heteromorphic translations is chosen arbitrarily, but it still serves the purpose of giving a rough estimation of how "free" or "paraphrastic" (Ingo 1991) a translation is.

It should be noted that not every possible structural aspect of the translations is considered in the analysis. Changes concerning function words such as conjunctions, prepositions and articles, or the number and definiteness of nouns, have been ignored in all cases except when they influence the classification clause function or type of phrase.

A more weighted measure of structural correspondence, a so-called *closeness measure*, in a sample can be arrived at with the following formula:

Closeness measure = (I+SI/2)/(I+SI+H)

where *I*, as above, is the number of isomorphic pairs, *SI* the number of semi-isomorphic pairs and *H* is the number of hetero-morphic pairs. The measure is designed so that it returns the value 1 if all sentence pairs are isomorphic, 0.5 if all pairs are semi-isomorphic and 0 if all pairs are hetero-morphic.

The classification of the semantic correspondence between source and target sentences is made in a similar fashion. Pairs of source and target sentences are classified as belonging to different categories depending on correspondences between their primary segments. In this classification four major categories are used:

EQ: Source and target sentences are regarded as having the same meaning;

LSP: The target sentence provides less information (is less specific) than the source;

MSP: The target sentence provides more information (is more specific) than the source;

OTH: Source and target sentences have some other relation to each other than the above.

For the first category, EQ, a similar rule of thumb as for the structural classification is used: If a word-for-word Swedish translation that gives a stronger semantic correspondence than the actual translation cannot be found, the translation pair is regarded as EQ.

It should be noted that the classification of semantic correspondence concerns the linguistically expressed meaning and not the interpretation of the content in context.

The structural analysis of each translation pair is complemented with a comparison of the corresponding translation segments' meanings. Cases of non-synonymy fall into three categories:

(i) More specific nucleus,

(ii) Less specific nucleus,

(iii) Nucleus with different meaning.

This classification is to a large part made on subjective grounds by the two evaluators. If there was disagreement among the evaluators, a bilingual dictionary was consulted (Norstedts Stora Engelsk-Svenska ordbok (Petti et al. 1994)). Two words were classified as equivalent if they were regarded as equivalents in the dictionary.

A translation pair is placed under the MSP category if the only changes in the translation are additions or the occurrence of more specific lexical items. Consequently, a pair is categorised as LSP if the only changes are deletions or the occurrence of less specific lexical items in the target. If there is mixture of such changes (both MSP and LSP contributing changes), or if simple functional changes give rise to semantic effects, the translation pair is categorised as OTH. A paraphrase is often meaning preserving, but in exceptional cases it can cause a change of meaning and lead to a different categorisation than EQ.

The classification of the translation pairs in the four semantic categories form the basis for the following measures of semantic correspondence:

Semantically equivalent translations:
EQ/(EQ+MSP+LSP+OTH)

Target specification degree:
(MSP - LSP)/(EQ+MSP+LSP+OTH)

Semantic degree of change:
(MSP+LSP+OTH)/(EQ+MSP+LSP+OTH)

10.3 Selection of texts

The sample translations incorporated in the tagged translation corpus come from the translation database described in the earlier chapters. From each of the eight texts 100 translation pairs were sampled randomly.

10.4 Tagging method

The tagged translation corpus was created in SGML by designing a Document Type Definition (DTD) which contained all tags and attributes described previously. This DTD was then used in conjunction with an SGML tagger called Author/Editor to mark up the corpus. All tags and attributes have been inserted manually.

The tagging process consisted of several steps:

1. Marking up functional segments in the source language and target language respectively.

2. Marking up the contrastive information in each translation pair. Here the focus was to record the significant changes from source to translation.

3. Verifying the tags and attributes.

To give the reader of an indication of what the mark-up looks like, here is an example from one of the software manuals:

```
<PAIR MAP="1-1" CONTENT="EQ" S-CORR="1">
<SOURCE><MAIN TENSE="PRESENT" MOOD="IMP" PRIMSEG="3"
TRANSEG="4"><TA>Then</TA> <FP>follow</FP><DO> the instructions
<PMOD>in the PivotTable Wizard</PMOD></DO>.</MAIN></SOURCE>

<TARGET><MAIN TENSE="PRESENT" MOOD="IMP" PRIMSEG="3"
TRANSEG="4"><FP>Följ</FP> <TA>sedan</TA> <DO>instruktionerna <PMOD>i
Pivottabellguiden</PMOD></DO>. </MAIN></TARGET></PAIR>
```

Figure 24 Example of SGML mark-up of a translation pair

This translation pair is analysed as being isomorphic (S-CORR=1) and as being semantically equivalent (EQ). The word order shift (then–sedan) present in the translation is regarded as a necessary grammatical shift and is therefore not recorded in the PAIR element. Both the source and target have the same mood (imperative) and have the same number of primary and translation segments (the definition and distinctions between these two will be explained in detail in section 10.7.

In the next sections I will present the tags and attributes used in the marked-up corpus.

10.5 Translation pairs – top level elements and attributes

The translation pairs have a top element, <PAIR>, which is divided into two subelements: <source> and <target>:

<PAIR>: <source>, <target>

The features of the whole translation pair are recorded in the PAIR element as SGML attributes. All translation pairs have attributes for the mapping relation (for example 1-1), semantic correspondence (EQ, MSP, LSP or OTHER), structural correspondence (the attribute is called S-CORR and takes either 1, 2 or 3 as values corresponding to isomorphic, semi-isomorphic and hetero-morphic respectively as outlined earlier). Also there is a list of all the recorded translation shifts that have been identified in the translation pair, e.g., mood shifts, level shifts, transpositions, etc.

The structure of the source and target elements is analogous. They can contain main clauses and possibly also co-ordinated main clauses.

Table 73. SGML elements and attributes used in the mark-up of structural and clausal segments

Clause/structural segments		
Tag	**Description**	**Attributes**
<PAIR>	A pair of source and target sentences	map, content, s-corr and translation operations
<SOURCE>	Source language sentence(s)	lang, prim-seg, tran-seg
<TARGET>	Target language sentence(s)	lang, prim-seg, tran-seg
<MAIN>	Main clause	voice, tense, mood
<SUB>	Subordinate clause	
<IPSUB>	Subordinate clause with infinitival construction	

The possible attributes in <source> and <target> are the same: language, number of primary and translation segments. For each main clause there are attributes to describe whether a sentence is active or passive and its tense and mood (using the attribute voice, tense and mood).

The SUB and IPSUB tags are not included directly under MAIN. Instead they occur as a syntactic descriptor to distinguish clauses from non-clauses. The information provided on the top level (from PAIR down to MAIN) can be described in the following diagram:

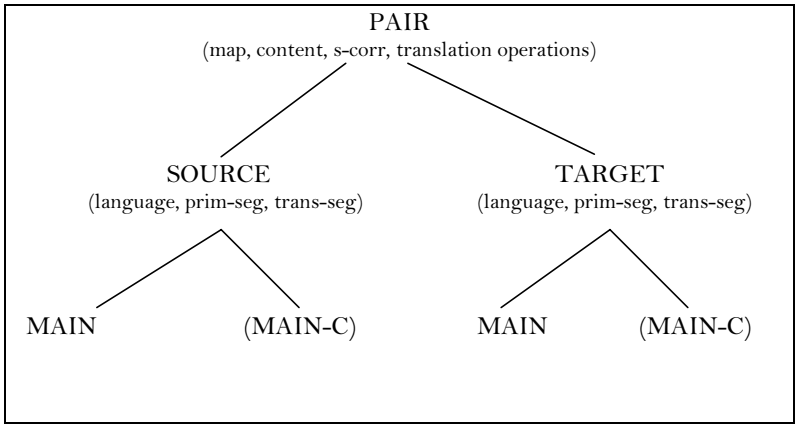


Figure 25 The top-level structure of the translation pairs

This means that all sentences described in the tagged corpora are regarded as full text sentences that will have at least one main clause included in the source and target elements.

10.6 Structural and functional elements

In main clauses there are two different kinds of units: *primary segments* and *translation segments*. The primary segments are defined as the main constituents in a main clause that carry a sentential function, such as subject, predicate verb,

object and adverbial. Translation segments are defined as the most specific units that carry content and are centred round a content word (nucleus). A primary segment can consist of a single translation segment if there is only one nucleus included in it, but it can also contain a number of translation segments, for example, in the source sentence in Figure 24 there are three primary segments: "then" (TA), "follow" (FP) and "the instructions in the PivotTable Wizard" (DO). The first two primary segments are also regarded as translation segments, but the last primary segment contains two translation segments, "the instructions" and "in the PivotTable Wizard". Function words (conjunctions, subjunctions, prepositions, particles, determiners and possessive pronouns) are included in their respective segments and are not marked up individually. They are used to determine the function of a particular segment in the clause.

The tags used to describe the primary segments are listed in the table below. Note that all these elements can also serve as translation segments if they are included in another primary segment, such as a subject in a temporal adverbial subordinate clause.

Table 74 Tags used to describe primary segments

Primary segments	
<SB>	Subject, in main or subordinate clause.
<FP>	Finite predicate verb.
<IP>	Infinite predicate verb.
<DO>	Direct or indirect object.
<PRED>	Predicative complement
<XO>	Clause complement.
<PO>	Prepositional object
<APP>	Apposition
<TA>	Temporal adverbial
<MA>	Manner adverbial
<PA>	Predicative adverbial
<LA>	Locational adverbial
<AA>	Other adverbial, such as conditional, causative, final or purposive.
<SA>	Sentential adverbial.

Some segments can however only occur as translation segments and not as primary ones. The following table lists these translation segments:

Table 75 Tags used to describe translation segments

Translation segments included in primary segments	
<NMOD>	Nominal modifier
<PMOD>	Prepositional modifier
<GMOD>	Genitive modifier
<XMOD>	Postpositional clausal modifier
<AMOD>	Adjectival modifier
<NUM>	Numeral modifier
<GA>	Grade adverbial

We are now ready to look at a more complex example of a translation pair. Consider the following example:

```
<PAIR MAP="1-1" CONTENT="EQ" S-CORR="3" IMPDECL="1" VERBINS="1"
REDDEL="1">
<SOURCE><MAIN VOICE="ACTIVE" TENSE="PRESENT" MOOD="IMP"
PRIMSEG="6" TRANSEG="13">
<AA><SUB>If <SB>you</SB> <FP>use</FP> <DO>a <NMOD>virus
protection</NMOD> program</DO> <LA>on your computer,
</LA></SUB></AA><FP>override</FP> <DO>it</DO><FP-C> or turn
<DO>it</DO> off</FP-C><TA><SUB> before <SB>you</SB> <FP>run</FP>
<DO>the <NMOD>Microsoft Excel</NMOD> Setup
program.</DO></SUB></TA></MAIN></SOURCE>
<TARGET><MAIN VOICE="ACTIVE" TENSE="PRESENT" MOOD="DECL"
PRIMSEG="4" TRANSEG="14"><AA><SUB>Om <SB>du</SB>
<FP>använder</FP><DO> ett program <PMOD>för
virusskydd</PMOD></DO><LA> på datorn</LA></SUB></AA><FP> bör</FP>
<SB>du</SB> <XO><IPSUB><IP>koppla bort</IP> <IP-C>eller stänga av</IP-C>
<DO>det</DO><TA><SUB> innan <SB>du</SB> <FP>kör</FP><DO>
installationsprogrammet <PMOD>för Microsoft
Excel.</PMOD></DO></SUB></TA></IPSUB></XO>
</MAIN></TARGET>
```

Figure 26 Example 2 of SGML mark-up of a translation pair

The source sentence consists of six primary segments (AA, FP DO, FP-C, DO and TA) with a total of thirteen translation segments (*you*, *use*, *virus protection*, *program*, *on your computer*, *override*, *it*, *or turn off*, *it*, *you*, *run*, *Microsoft Excel*, *Setup program*). The target sentence contains four primary segments (AA, FP, SB, XO) and fourteen translation segments *(du*, *använder*, *ett program*, *för virusskydd*, *på datorn*, *bör*, *du*, *koppla bort*, *eller stänga av*, *det*, *du*, *kör*, *installationsprogrammet*, *för Microsoft Excel*). The translation operations registered here are that the imperative source sentence has turned into a declarative target sentence (IMPDECL), a modal verb has been inserted (*bör*) and a repeated direct object (*it*) has been deleted in the target. The pair has been classified as being semantically equivalent (CONTENT=EQ) and heteromorphic (S-CORR="3") for reasons given in sections 10.7.2 and 10.7.3.

10.7 Tags and attributes for translation shifts

10.7.1 Attributes for recording changes between source and target sentences

The following is a list of the attributes and values used in the tagging of correspondences. The attributes are divided into complex operations which involve multiple segments, category shifts, transpositions, lexical shifts and non-1-1 operations (deletion, addition, convergence and divergence).

10.7.1.1 Complex operations (involving multiple segments)

VOICE SHIFT

a) Active-->Passive

This operation involves changing the voice of the target sentence into passive (compared to active voice in the source). It also entails the category shifts necessary for the passive operation, namely to change the subject of the source to a prepositional object (agentive or instrumental) in the target as well as turning the object of the source into the subject of the target.

Attribute: ActPass
Value: <num>
Entails: SB-->PO, DO/PO-->SB
Example:
Paradox stores *important information about a table's primary key in an index (.PX) file.*

I Paradox sparas *viktig information om tabellens primärnyckel i en indexfil (.PX).*

b) Passive-->Active

This is the inverse operation of a) described above, where a passive source sentence is changed into an active target sentence.

Attribute: PassAct
Value: <num>
Entails: PO-->SB, SB-->DO/PO
Example:
They ***were grabbed*** *by Arab workmen, picked up by the fleece, and thrown writhing into the truck while everyone shouted curses.*

Arabiska arbetare ***högg tag*** *i dem, de lyftes upp i ullen och kastades vrenskande in i trucken medan alla ropade förbannelser.*

MOOD SHIFT

The shifts of sentence moods present in the corpus involve changing declaratives to imperatives or vice versa, as well as changing questions into declaratives and vice versa. The first two are by far the most frequent ones and therefore these are given separate descriptions.

a) Imp-->Decl

The imperative to declarative mood shift entails that a subject is added in the Swedish target

Attribute: ImpDecl
Value: <num>
Entails: ExpAdd: SB
Example:
If you want to check the query, ***choose Datasheet*** *from the View menu.*

Om du vill kontrollera frågan, ***väljer du Datablad*** *från Visa-menyn.*

b) Decl-->Imp

The declarative to imperative mood shift entails the inverse from a) above, namely that a target subject is deleted.

Attribute: DeclImp
Value: <num>
Entails: RedDel:SB
Example:
You can back up *all programs and data that you want to save and then format the hard disk partition when you install OS/2 2.1.*

Ta backup *på allt du vill behålla, och formatera därefter hårddisken vid installationen av OS/2 2.1.*

c) Other mood shift

This category involves all other mood shifts than the ones mentioned in a) and b), for example when questions are changed into declaratives or imperatives.

Attribute: XMood
Value: <num>

LEVEL SHIFT (lowering or raising)

In this category the hierarchical structure of a sentence has changed, by for example deleting or inserting modal verbs, or nominalizations of verb phrases.

a. Verb deletion[22]

A modal verb in the source is deleted in the target. This will often increase the number of primary segments in the target, because a complex verb phrase will be split up into a flatter structure, and objects and adjuncts will be raised one level in the sentential structure.

Attribute: VerbDel
Value: <num>
Entails: RedDel: FP/IP, ExpDiv: XO-->FP, DO/PO/adverbials/PRED...)

[22] Could be either finitival or infinitival verbs, e.g. "To be able to do that..." - "För att göra detta..." or "You can write this..." - "Du skriver detta..."

Example:
***Plan** to base the main form on the Customers table and the subform on the Quarterly Orders by Product query.*

Basera huvudformuläret på tabellen Kunder och underformuläret på frågan Order per kvartal och produkt.

b) Verb insertion

The inverse of a) above. This will often reduce the number of primary segments in the target as the main verb phrase gets more complex.

Attribute: VerbIns
Value: <num>
Entails: ExpAdd: FP/IP, RedCon: (IP, DO/PO/adverbials/PRED...)-->XO
Example:
All MAPI users have write access to their Office Vision/400 folders on the AS/400.

*Alla MAPI-användare **ska** ha skrivbehörighet till sina Office Vision-mappar i AS/400.*

c) Non-clause to clause

In this level shift a non-clause element (such as a prepositional, adverbial or nominal modifier) is translated by a subordinate clause.

Attribute: NCC
Value: <num>
Entails: ExpDiv: non-clausal segment-->SUB/IPSUB)
Example:
*Bray and Vivien speculated about **the celebrations in the African villages and townships**.*

*Bray och Vivien spekulerade **om hur man firade den stora händelsen i de afrikanska städerna och byarna**.*

d) Clause to non-clause

Here a clause element (subordinate or main clause) is translated by a non-clause element (for example, with a noun phrase or a prepositional phrase), which entails a convergence operation.

Attribute: CNC
Value: <num>
Entails: RedCon: IPSUB/SUB-->non-clausal segment

Example:
But suddenly the music stops ***and a terrorist bomb is reported.***

Men plötsligt tystnar musiken ***för en rapport om en terrorbomb****.*

e) Infinitival to finitival verb clause

A common operation in the translation from English into Swedish is to change a verb clause with an infinitival verb into a full subordinate clause. This will entail that the subject is made explicit and that the verb changes to a finitival verb in the target.

Attribute: FinCl
Value: <num>
Entails: ExpAdd: SB, CatShift: IP-FP
Example:
Use the following procedure ***to edit either an embedded*** *or a linked object in a bound object frame.*

Använd följande procedur ***om du vill redigera ett inbäddat*** *eller länkat objekt i ett bundet objektfält.*

f) Finite to infinite verb clause

This shift is the inverse operation of e) above and consequently demands a suppression of the subject in the target as well as a change of the finiteness of the verb.

Attribute: InfCl
Value: <num>
Entails: RedDel: SB, CShift: FP-IP
Example:
He ordered the beer that the other said ***he would have****.*

Han beställde en öl, som mannen sade sig ***föredra****.*

g) Main to sub-clause (co-ordinated main clause or verb phrase to sub-clause)

A co-ordinated main clause or verb phrase can also be lowered in the target, that is, translated by a subordinate clause in the target.

Attribute: SubCl
Value: <num>
Example:
She was level with Bray ***and he half-rose politely.***

Just då var hon ungefär i höjd med Bray ***som artigt reste sig till hälften.***

h) Sub to main clause (sub clause to co-ordinated main or verb clause)

This is the inverse operation of g) above.

Attribute: MainCl
Value: <num>
Example:
Stephen Wentz smiled and showed off a little ***as he put down a bottle of brandy and two fancy goblets****.*

Stephen Wentz log och ***satte en konjaksflaska och två aromglas på bordet*** *med en lite extra yvig gest.*

10.7.1.2 Paraphrases

Paraphrases come in two kinds: *complex* and *partial* paraphrases. Both types of paraphrases occur when a set of translation segments in the target is mapped to a structurally different set of translation segments in the target. The resulting translation is too complex to divide into different suboperations. The difference between a complex and a partial paraphrase is that the former has scope over the whole sentence, whereas the latter operates on a part of the sentence. Note that the paraphrase operation does not entail that content is preserved, i.e., a translation can therefore be considered to have any of the possible attributes for CONTENT. This means that a paraphrase can result in a target translation that is equal, more specific, less specific or different concerning CONTENT.

Attribute: ComplPar
Value: <num>
Example:
Pressing the Cursor Blink (CrBnk) key while pressing and holding the Alt key turns on and off the blinking of the cursor.

Om du håller ned Alt-tangenten och trycker på den här tangenten börjar eller slutar markören blinka.

Attribute: PartPar
Value: <num>
Example:
If you are installing on a computer that does not contain an operating system, ***you will probably want to follow the basic installation procedure****.*

Om du installerar på en dator som inte tidigare har något operativsystem ***är grundinstallationen enklast****.*

10.7.1.3 Category shifts

Three different kinds of category shifts are handled: simple shifts, shifts which involve the change of the categorisation of a subordinate clause, often in terms of shifting one type of adverbial to another, and a third type involving negation.

CATEGORY SHIFTS (1-1)

a) Simple category shifts

Attribute: CShift
Value: <num>
Example:
In Form view, you can usually see all the fields ***for one record*** *at a time.*

I ett formulär kan du vanligtvis titta på ***en posts*** *alla fält samtidigt.*

The above is an example of when a prepositional modifier has been shifted into a genitival modifier.

b) Clausal category shifts

Attribute: ClCShift
Value: <num>
Example:
On printed forms, you use a page break ***to start a new page within a section****.*

I utskrivna formulär kan du använda sidbrytning ***när du påbörjar en ny sida i mitten av ett avsnitt.***

This example illustrates the translation of a purposive adverbial in the source into a temporal adverbial in the target.

c) Negation shifts

Attribute: NegShift
Value: <num>
Example:
… that there is ***no equivalent mouse function****.*

… att det ***inte finns någon motsvarande musfunktion****.*[23]

10.7.1.4 Transpositions

Two different kinds of transpositions are described in the model. The first is called topicalization and involves moving a primary or translation segment into the initial position in the target. The second transposition is used for all non-fronting transpositions.

[23] This translation is regarded as having a voluntary negation shift. The source sentence could have been translated more directly by *"ingen motsvarande musfunktion"*. See page 169 for an example of a necessary negation shift.

a) Topicalization

Attribute: TranspT
Value: <num>
Example:
*You also need up to an additional 5 MB of **hard disk** space.*

***På hårddisken** behöver du 5 MB extra utrymme.*

b) Other transposition

Attribute: TranspX
Value: <num>
Example:
*Adjust the width of the datasheet columns, **if necessary**, so they fit in the subform control.*

*Justera **eventuellt** bredden av databladets kolumner så att de passar i underformulärets kontroll.*

10.7.1.5 Non-1-1 operations

The non-1-1 operations contain operations where there is no 1-1 correspondence between translations segments in the source and target. These operations include additions, deletions, divergences and convergences.

a) Additions

Attribute: ExpAdd
A translation segment is added in the target.
Value: <num>
Example:
Datasheets are the easiest type of subforms to create.

*Datablad är den **snabbaste** och enklaste typen av underformulär.*

b) Divergences

A source translation segment is distributed over two target translation segments.

Attribute: ExpDiv
Value: <num>
Example:
*Change **how** data is calculated.*

*Ändra **det sätt på vilket** informationen beräknas.*

c) Deletions

A source translation segment is deleted in the target.

Attribute: RedDel
Value: <num>
Example:
Datasheets are the easiest type of subforms ***to create****.*

Datablad är den snabbaste och enklaste typen av underformulär.

d) Convergences

Two (or more) source translation segments correspond to one target translation segment.

Attribute: RedCon
Value: <num>
Example:
In this exercise, the location of the branch office ***is variable****.*

I den här övningen ***varierar*** *filialkontorets adress.*

10.7.1.6 Lexical shifts

Lexical shifts involve changing the meaning of translation segment from the source to the target. There are three different kinds of lexical shifts, (a) a shift where a source translation segment is translated with a more specific lexical item, (b) a shift where a source translation segment is translated with a less specific (more general) lexical item, and (c) a shift where the translation is neither more or less specific or equivalent to the source lexical item (a different meaning).

a) "More specific" shift

Attribute: LexM
Value: <num>
Example:
Returning to ***it*** *next day, I found Faulkner guilty of no offense.*

När jag återgick till ***texten*** *nästa dag fann jag Faulkner oskyldig till något klandervärt.*

b) "Less specific" shift

Attribute: LexL
Value: <num>
Example:
Follow the directions ***in the Form Wizard dialog boxes****.*

Följ instruktionerna ***i dialogrutorna****.*

c) Other lexical shift

Attribute: LexU
Value: <num>
Example:
*The scale line disappears and the setting is **completed**.*

*Skalan försvinner och inställningarna **sparas**.*

10.7.2 The attribute CONTENT and some considerations

The general idea is to record the explicit semantic relationship between the source and the target. On a pragmatic level the sentences may be equal, but the focus is on what is expressed explicitly in the source and target manifestations, which means that it is what and how translation segments are manifested that govern the value of the attribute CONTENT. Below a list of criteria for deciding the value of CONTENT is shown:

1. A definite description in the target instead of pronoun makes the translation MSPEC and vice versa.

2. A deleted translation segment in the target makes the translation LSPEC.

3. An added translation segment in the source makes the translation MSPEC.

4. A translation segment tagged as LEXM makes the translation MSPEC (and a LEXL tag makes it LSPEC).

5. If there is more than one change which influences the CONTENT attribute, the translation is considered OTHER if these changes have different directions (for example, one LEXL and one LEXM).

6. The operation FINCL does not cause any change in the CONTENT value (that is a change from SUB to IPSUB does not influence CONTENT if the implicit subject is the same as in the main clause).

7. The operation IMPDECL when together with VERBINS (modal) does not cause any change in CONTENT. (For example *Open the document—Du bör öppna dokumentet.*

8. The operation IMPDECL without VERBINS does cause change in CONTENT towards OTH(ER). (For example *Open the document—Du öppnar dokumentet.*

9. Differences in conjunctions (and, or, but) are not recorded as CONTENT shifts as they are not tagged as translation segments.

10. A CLCSHIFT from AA (conditional, causative, purposive adverbial) to TA (temporal adverbial) causes CONTENT to be LSPEC.

11. A "to-clause" (infinitival clause with purposive effect) which is translated as a conditional clause causes CONTENT to be OTH(ER). Thus, such a shift is considered to change the meaning.

12. A CLCSHIFT from TA (temporal adverbial) to AA (conditional, causative, final adverbial) causes CONTENT to be MSPEC.

13. CSHIFTs in modifiers such as GMOD or NMOD to PMOD is not recorded as a CONTENT shift.

14. Co-ordinated verb phrases with identical objects in the source where one of the objects is deleted in the target are not recorded as changing the CONTENT (for example, *open it and close it* vs. *open and close it*). The opposite change is handled the same way, that is, when an identical object in a Coordinated verb phrase is added in the target.

15. A VERBINS or VERBDEL operation causes CONTENT to be OTH(ER) (if the inserted or deleted verb is not elliptical).

10.7.3 The attribute S-CORR and some considerations

In section 10.2.2 the general rules of thumb to determine whether a translation is to be regarded as isomorphic, semi-isomorphic or heteromorphic were described. In the SGML mark-up these characteristics are recorded in the S-CORR attribute under the PAIR element with numerical values (1, 2 and 3). The value 1 stands for an isomorphic relation, 2 for semi-isomorphic and 3 for a heteromorphic relation. The list below contains the more detailed rules that were applied when deciding to what category a translation belonged:

1. If there is a complex paraphrase operation (COMPLPAR), the S-CORR attribute is always set to heteromorphic (3).

2. Lexical shifts (LEXM, LEXL and LEXU) do not influence the value of S-CORR.

3. If there are any structural changes the S-CORR value is always 2 or 3 (semi-isomorphic or heteromorphic).

4. If there are less than 8 translation segments in the source, 2 or more distinct structural operations cause S-CORR to be 3 (heteromorphic).

5. If there are at least 8 translation segments in the source, 3 or more distinct structural operations cause S-CORR to be 3 (heteromorphic).

10.8 Examples

Let us illustrate with two examples where the correspondence model has been applied. The previous example shown in Figure 26 earlier is shown again below:

```
<PAIR MAP="1-1" CONTENT="EQ" S-CORR="3" IMPDECL="1" VERBINS="1"
REDDEL="1">
<SOURCE><MAIN VOICE="ACTIVE" TENSE="PRESENT" MOOD="IMP"
PRIMSEG="6" TRANSEG="13">
<AA><SUB>If <SB>you</SB> <FP>use</FP> <DO>a <NMOD>virus
protection</NMOD> program</DO> <LA>on your computer,
</LA></SUB></AA><FP>override</FP> <DO>it</DO><FP-C> or turn
<DO>it</DO> off</FP-C><TA><SUB> before <SB>you</SB> <FP>run</FP>
<DO>the <NMOD>Microsoft Excel</NMOD> Setup
program.</DO></SUB></TA></MAIN></SOURCE>
<TARGET><MAIN VOICE="ACTIVE" TENSE="PRESENT" MOOD="DECL"
PRIMSEG="4" TRANSEG="14"><AA><SUB>Om <SB>du</SB>
<FP>använder</FP><DO> ett program <PMOD>för
virusskydd</PMOD></DO><LA> på datorn</LA></SUB></AA><FP>
bör</FP> <SB>du</SB> <XO><IPSUB><IP>koppla bort</IP> <IP-C>eller
stänga av</IP-C> <DO>det</DO><TA><SUB> innan <SB>du</SB>
<FP>kör</FP><DO> installationsprogrammet <PMOD>för Microsoft
Excel.</PMOD></DO></SUB></TA></IPSUB></XO>
</MAIN></TARGET>
```

Figure 27 Example 2 of SGML mark-up of a translation pair

The CONTENT attribute is here set to EQ and the S-CORR attribute to "3" (heteromorphic). At a first glance this may seem surprising as there are three distinct shifts recorded (IMPDECL, VERBINS and REDDEL). In the target the mood has changed to deccarative, but this is accompanied by a modal verb insertion which is considered to be semantically equivalent to the target imperative mood (see item 7 in section 10.7.2). The deleted source translation segment is the first "it" in the co-ordinated verb phrase ("override it or turn it off"). This has been translated by "koppla bort eller stänga av det" where the first direct object is missing. As item 14 in section 10.7.2 states, this is a case of when the CONTENT attribute should not be changed, as it does not affect the meaning of the co-ordinated phrase.

Another example is shown below:

```
<PAIR MAP="1-1" CONTENT="MSPEC" S-CORR="3" PASSACT="1"
EXPADD="1">
<SOURCE><MAIN TENSE="PRESENT" MOOD="IMP" PRIMSEG="3"
TRANSEG="5">
<FP>Release</FP><DO>the mouse button</DO><TA><SUB> when<SB>
the <AMOD>desired</AMOD> level</SB><FP> is
reached</FP></SUB></TA>.</MAIN></SOURCE>
<TARGET><MAIN TENSE="PRESENT" MOOD="IMP"
PRIMSEG="3" TRANSEG="6"><FP>Släpp</FP>
<DO>musknappen</DO><TA><SUB> när <SB>du</SB>
<IP>nått</IP> <DO><AMOD>önskad</AMOD>
nivå</DO></SUB></TA>. </MAIN></TARGET>
</PAIR>
```

Figure 28 Example 3

The translation pair is considered to be heteromorphic because there are two structural shifts (PASSACT and EXPADD) and only five translation segments in the source. The passive-to-active shift is present in the temporal adverbial subordinate clause and it is in that clause that the addition also has taken place, namely the addition of an explicit agent (*du*). The addition of the explicit agent causes the target to be more explicit than the source, which accounts for the MSPEC value of the CONTENT attribute.

10.9 Summary

In this chapter a model for describing structural and semantic changes between a source text and target text has been presented. By focusing on the voluntary structural and semantic shifts, it is possible to describe the degree of change between a source and target sentence. Thus, it also becomes possible to provide detailed descriptions and measures of how "free" or "close" a given translation is in relation to its source.

The correspondence model has been applied to a sample of the Linköping Translation Corpus. The results of the correspondence analysis is reported in the next chapter.

11 Translation correspondence – an analysis

In this chapter the correspondence model outlined in the previous chapter is applied to the tagged translation corpus. The analysis has been made on 100 randomly sampled sentence translations from each of the eight texts in the Linköping Translation Corpus. The eight texts can also be divided into four different groups depending on text type and what kind of translation method that was used during the translation.

A. **Computer manuals translated manually** (without any software tools)

1) Microsoft Access User's Guide version 2.0

2) Microsoft Excel User's Guide version 5.0

B. **Computer manuals translated with translation memories**

3) IBM OS2 Installation Guide and User's Handbook (version 2.1)

4) IBM InfoWindows User's Handbook

5) IBM Client Access for Windows User's Guide

C. **Fiction**

6) Nadine Gordimer's "Guest of Honour"

7) Saul Bellow's "To Jerusalem and Back"

D. **Automatic translation**

8) The ATIS dialogues

The texts have been tagged according to the correspondence model, in the SGML format described in chapter 1. The tagging was first performed by the author and then verified by Lars Ahrenberg. It must be pointed out that the correspondence model was extended and revised several times due to experiences from the actual tagging. Therefore, the tag set changed during the course of the project until it took the final form that is described in chapter 10. After the tag set was stable, the texts were checked in a final round of verifications by both Lars Ahrenberg and me.

11.1 Hypotheses

Before the analysis of the translations began, there were a number of expectations that I wanted to have verified or contradicted by the data:

1. *Translation memories should make translations structurally more equivalent compared to pure human translations.*

   This hypothesis stems from the fact that the target text can be reused in different contexts and that too drastic structural changes would make the reuse of translations more difficult.

2. *According to Baker (1993, 1995) there should be a higher degree of specification in the translations than in the source texts.*

   This would mean that there should be a higher number of MSPEC values for the CONTENT attribute than the other values (EQ, OTHER and LSPEC).

3. *Automatic translations should have a very high degree of structural and semantic correspondence.*

   This means that structural changes should be rare for this kind of translations and that meaning is preserved from source to target with small modifications.

4. *It should be possible to couple certain translation operations to certain text types.*

   In other words, some of the translation operations would be much more common in computer manual translations than in translations of fiction.

5. *The discovery of the apparently ineffecient use of translation memory-based translation tools, uncovered by the analysis tools in chapter 6, should also be visible in a closer structural and semantic correspondence.*

   In the recurrence and discrepancy analyses of the texts made earlier that were purely string-based (with no linguistic information at all), some unexpected results were observed regarding translations of manuals with and without translation memories. In two of the TM-translated IBM manuals, the data revealed an apparently inefficient use of the translation support tools, which seemed to result in more communicative and free translations. In the more detailed analysis of the tagged translation database, it is possible to examine if this communicative translation strategy also is manifested in terms of structural and semantic change. Also, if the third TM translation (Client) is "more efficient", in what way, if at all, can this be seen in the more detailed analysis which is the focus of this chapter?

11.2 Structural correspondences

The structural changes that have taken place in the translations are recorded as values of the S-CORR attribute. Translations can be either isomorphic (no changes), semi-isomorphic (minor changes) or heteromorphic (major changes) from a structural point of view.

The table below summarises the structural relationships between the source and target texts in the eight samples. Note that the figures in the table can also be taken as rough percentages as there are 100 samples in each text.

Table 76. Structural correspondence in the translations

	Manuals - Human translation		Manuals - TM translation			Fiction		Automatic translation
	ACCESS	EXCEL	OS2	INFOWIN	CLIENT	GORDI-MER	BELLOW	ATIS
Isomorphic	21	20	10	12	38	33	40	97
Semi-isomorphic	31	36	24	19	21	24	35	3
Hetero-morphic	48	44	66	69	41	43	25	0

Here the automatic characteristics of machine translation (ATIS) are apparent. Ninety-seven out of 100 sampled sentences have been translated in such a way that the structure of the source is directly transferred to the translation. For the other texts, it is more difficult to distinguish any obvious trends. Although the Access and Excel translations have similar values, they are more isomorphic than the TM-translated IBM texts, except for the Client translation. Let us take a closer look at the translations by applying the *closeness measure* described in the previous chapter.

Closeness measure = (I+SI/2)/(I+SI+H)

where *I* is the number of isomorphic pairs, *SI* the number of semi-isomorphic pairs and *H* is the number of heteromorphic pairs. The continuum from 0 to 1 indicates the extensiveness of the structural changes, ranging from complete paraphrasing in the whole sample (0) to no structural changes (1).

Table 77 Structural correspondence measured as structural closeness

	Manuals - Human translation		Manuals - TM translation			Fiction		Automatic translation
	ACCESS	EXCEL	OS2	INFOWIN	CLIENT	GORDI-MER	BELLOW	ATIS
Closeness measure-ment	0.365	0.38	0.22	0.215	0.485	0.45	0.575	0.985

The text that exhibits the highest closeness measure is ATIS, which can be regarded as completely isomorphic (0.985). The Microsoft translations show similar values for structural correspondence when compared to each other, but higher values than the first two TM-translated manuals from IBM. As was shown earlier, this is consistent with the data from the discrepancy analysis made earlier. The IBM translations, especially OS2 and INFOWIN, are more communicative than any one of the other translations. Without the discrepancy data and information from the translators, these figures would seem to be as surprising as the discrepancy data were when they were first uncovered; translation memory translations should, in theory, steer the translation towards higher structural correspondences, but the importance of the IBM translation culture is clearly visible in this type of analysis too. When the IBM translators changed their translation strategy, as in the example of CLIENT, the closeness

measure becomes twice as high compared to the earlier attempts with translation memory tools.

It is surprising though, that the more "artistic" texts, i.e. the two novels, are structurally more isomorphic than the computer manuals. In other words, the fiction translators have changed less of the source text structure than the translators of the computer manuals have. The reason for this can probably be found in the general difference between the text types. In translations of fiction the individual style of a writer is more important than in strictly informative and instructional text. The fiction writer's style should to a certain extent be transferred to and preserved in the translation of a novel, whereas this aspect is less important in the translation of technical documentation. In technical translations, correctness and readability is more important, which means that the translator can, and should, avoid being too restricted by the structure of the source text.

Automatic translation with the aid of machine translation tools is structure preserving. It is very difficult to design machine translation programs in such a way that they produce a "natural" degree of structural change. Automatic translation programs take the shortest path to a grammatical translation. The whole MT philosophy entails a very high level of structural correspondence. The figures for structural closeness above give us some interesting perspectives. Let us for a moment imagine that there existed automatic translation software that could translate both the computer manuals from Microsoft and IBM as well as the fiction, meaning that the machine translation system could cover both the syntax, semantics and lexicon for these texts. Furthermore, we would assume that these translations were made with the same approach as the translation of the ATIS text, that is, producing structural correspondence figures of around 95-99 per cent. The result of such a hypothetical comparison would be that fiction texts (especially the BELLOW text) are better candidates for machine translation than the computer manuals.

Structural correspondence is however only one side of the coin. Let us now turn to what has happened to content, i.e., the semantic change from source to target text.

11.3 Semantic correspondences

In the chapter 1 there is a detailed outline of how to judge what happens in a translation from a semantic point of view. Four values on the sentence level were used to characterise the semantic correspondence between a source and target sentence. These were EQ (semantically equivalent), MSPEC (the target sentence is semantically more specific than the source sentence), LSPEC (the target sentence is semantically less specific than the source sentence) and OTHER (the target sentence carries a different semantic content than the target sentence). The distribution of the values for semantic correspondence is shown in Table 78 below.

Table 78. Semantic correspondence

	Manuals - Human translation		Manuals - TM translation			Fiction		Automatic translation
	ACCESS	EXCEL	OS2	INFOWIN	CLIENT	GORDI-MER	BELLOW	ATIS
EQ	36	37	9	13	37	37	71	90
MSPEC	11	14	10	10	6	27	11	1
LSPEC	24	19	21	21	34	4	4	0
OTHER	29	30	60	56	23	32	14	9

By looking first at the ratio of EQ translations, we get a similar picture as in the structural analysis. The earliest IBM translations (OS2 and INFOWIN) show the lowest number of semantically preserving transfers. At the other end of the scale ATIS is placed with 90 per cent EQ transfers. What is more notable is a division between the fiction translations and the manual translations as far as specificity is concerned. In the novels, the translators have added information (i.e., there are more MSPEC translations than LSPEC translations) whereas the computer manual translators have removed information (more LSPEC translations than MSPEC translations). The non-EQ values in the ATIS translation (1 MSPEC and 9 OTHER) all stem from mistaken lexical transfers. For example, the verb *be* in the source is transferred as *finnas*, *that* as *den där* and acronyms as *BWI* are spelt out in the target in an unintended way. The lexical transfer mistakes are just side effects of the way that machine translation software works. In many other contexts these transfers are the intended ones, but not always.

A different perspective on the data in Table 78 can be taken by applying the three measures for semantic correspondence described in chapter 1:

- SED –Semantically Equivalent translations Degree:
  100xEQ/(EQ+MSP+LSP+OTH)
- SDC – Semantic Degree of Change:
  (MSP+LSP+OTH)/(EQ+MSP+LSP+OTH)
- TSD – Target Specification Degree:
  (MSP - LSP)/(EQ+MSP+LSP+OTH)

The SED measure is a simple measure of the ratio of maximal semantic correspondence in the sample. The second semantic measure is the complement to the SET measure and gives an indication of the degree of the total semantic change in the translation. The final measure, TSD, describes the direction of specificity, i.e., whether the translation is geared towards a higher or lower degree of specification. A summary of the figures for these measures applied to the translations is given in Table 79.

Table 79. Semantic change in the sample

	Manuals - Human translation		Manuals - TM translation			Fiction		Automatic translation
	ACCESS	EXCEL	OS2	INFOWIN	CLIENT	GORDI-MER	BELLOW	ATIS
SED	0.36	0.37	0.09	0.13	0.37	0.37	0.71	0.90
TSD	-0.13	-0.05	-0.11	-0.11	-0.28	0.23	0.07	0.01
SDC	0.64	0.63	0.91	0.87	0.63	0.63	0.29	0.10

What was perhaps difficult to distinguish in Table 78 is more obvious here. All the computer manual translations have negative values for TSD, which means that all these translations have a tendency to exclude information from the source text in the translations. In the fiction translations, explicit information has been added to the translations, which is shown by the positive values for TSD. The semantic degree of change (SDC) is highest in the first two IBM translations (0.87–0.91) and lowest in the ATIS and BELLOW translations. How should we interpret the translations judging from these figures? In my opinion the following conclusions can be drawn:

1. The fiction translations are characterised by the tendency that explicit information is added in the translations.

2. The translations of manuals from (American) English into Swedish are characterised by the tendency that explicit information is deleted in the translations.

3. There is no distinguishable effect of the use of translation memories as far as semantic correspondence is concerned. The human translations of computer manuals exhibit similar semantic correspondence values when compared to translations performed with the aid of translation memories.

4. The two fiction texts in the sample are very different as far as the degree of semantic change. Due to the limited number of texts, it is impossible to say anything specific about the translation of this text type.

5. Similar to what the data from the structural correspondence showed, the BELLOW text is the translation that is closest to meaning preservation of all the texts (71 per cent SED). The selection of lexical items is naturally much more complicated in this text type, but the data imply that the BELLOW translation, from a strict semantic preservation point of view, would be the easiest to recreate given the appropriate machine translation software and disregarding the increased problems of lexical choice.

6. If we aimed for the translation characteristics present in the first IBM translations (made with TM tools) with the aid of automatic translation, we would encounter grave difficulties. As the semantic changes are so dominating in the OS2 and INFOWIN translations

(around 90 per cent SDC), such translations would be virtually impossible for any foreseeable MT system to accomplish.

7. Automatic translations (ATIS) preserve meaning. Apart from the mistakes in the transfer of lexical items, the semantic correspondence is 100 per cent.

Given the data from the analyses of both structural and the semantic correspondence, a comparison was made to see if there were any groups of translations that exhibited similar, significant changes. To do this, a chi^2 test was used, where the input data were the numerical values presented in Table 76 and Table 78. The chi^2 test was used to investigate whether the differences in structural and semantic correspondence were significantly non-random. The result turned out to be that all possible pairs of translations were significantly different, except two of the pairs, namely the two translations from Microsoft (EXCEL and ACCESS) and the first two translations from IBM (OS2 and INFOWIN). Company guidelines and the specific text type may be the reason that these two texts show some similar characteristics, even though different translators worked on the texts.

In the next section the different structural and semantic operations/shifts are investigated in more detail.

11.4 Focus on change

In the mark-up of the translation samples 28 different types of translation operations were recorded. Instead of listing the number of each operation for every text, certain operations can be clustered into groups for a better overview (see Table 80 below). A more detailed account of the specific operations can be found in the chapter 1.

Table 80. Groups of translation operations

Group	Specific operations/shifts	Description
Voice operations	**ACTPASS, PASSACT, XVOICE**	Changes in sentence voice
Mood operations	**IMPDECL, DECLIMP, XMOOD**	Changes in sentence mood
Level shifts	**VERBINS, VERBDEL, NCC, CNC, FINCL, INFCL, MAINCL, SUBCL**	Level shifts on a verbal or sentential clause, involves changes to the hierarchical structure of a sentence
Paraphrases	**GLOBPAR, PARTPAR**	Global or partial paraphrases
Category shifts	**CSHIFT, CLCSHIFT, NEGSHIFT**	Shift of functional category (either on translation segment or whole clause)
Addition/deletions	**EXPADD, EXPDIV, REDDEL, REDCON**	Additions, divergences, deletions and convergences of translation segments
Transpositions	**TRANSPT, TRANSPX**	Transpositions into topical or other position
Lexical operations	**LEXM, LEXL, LEXU**	More specific, less specific or "semantically different" lexical transfer

In Table 81 below, the number of translation operations for each text is presented. Here the exact figures are not so important; instead it is the presence and proportion of certain operations that are focussed.

Table 81. Number of translation operations in each translation

	Manuals - Human translation		**Manuals - TM translation**			**Fiction**		**Automatic translation**
	ACCESS	**EXCEL**	**OS2**	**INFOWIN**	**CLIENT**	**GORDI-MER**	**BELLOW**	**ATIS**
Voice op.	27	18	27	38	17	8	5	0
Mood op.	6	15	16	10	9	0	0	0
Level shifts	55	58	56	40	32	31	24	0
Para-phrases	11	11	35	28	16	31	12	0
Cat. shifts	36	25	14	24	13	11	14	0
Add/del	75	66	110	74	37	49	27	3
Trans-positions	13	10	9	11	12	9	11	0
Lexical op.	40	41	85	70	35	31	19	10
TOTAL	**263**	**244**	**352**	**295**	**171**	**170**	**112**	**13**

There are certain interesting observations to be made from the data in Table 81. The first aspect regards the difference between the fiction and computer manual

translations. No changes of sentence mood occur in either of the two novels. In the computer manuals between 6 and 16 per cent of all translations change sentence mood, usually from the declarative to the imperative, or vice versa. (This will be investigated in more detail later on in this chapter.) The second observation is the question of voice changes. Voice changes occur in all translations except ATIS, but they are considerably more frequent in the computer manuals compared to the fiction translations. The third observation is a natural implication from the data presented earlier: the machine translated ATIS text exhibit very few operations and is therefore, once again the easiest to distinguish. The visible changes that have taken place in the ATIS text are direct results of how the syntactic or lexical transfers are designed within the MT system. A fourth observation is that the total number of translation operations recorded reflects the characteristics observed with the discrepancy analysis made in earlier chapters. The relative differences between the two "early" TM translations at IBM (OS2 and InfoWin) on the one hand and the later TM translation (Client) on the other hand, where the new technology had been more accepted by the translators are shown by the fact that the total number of translation operations have been reduced to roughly half compared to the OS2 and Client translations.

Before I turn to the question of mood and voice changes I would just briefly like to go through each of these group of operations and make some general comments:

11.4.1 Level shifts (clause operations)

A level shift causes the hierarchical structure of a sentence to change, for example, by deleting or inserting modal verbs, or by nominalizations of verb phrases. The insertion of a modal verb in the target is often done in conjunction with a change in sentence mood. For example, in EXCEL, there are fifteen changes from the imperative to the declarative, and in nine of these changes a modal verb is inserted. The most frequently used operations overall are in descending order FINCL (infinitival clause-to-finite clause), CNC (clause-to-non-clause), VERBINS (insertion of modal verb) and NCC (non-clause-to-clause).

11.4.2 Paraphrases

The Microsoft translations contain the least number of paraphrases in the corpus, not counting ATIS. The more communicative translations, to which OS2 belongs, INFOWIN and GORDIMER, have around three times as many paraphrases as the ACCESS and EXCEL translations. If we look at the distribution of global and partial paraphrases (that is, whether the paraphrase spans the whole sentence or only a portion of it), the fact is that there are proportionally more global paraphrases in the manuals than in the fiction translation. In other words, it is more frequent in the manuals to rephrase a whole sentence than it is in the novels. In for example GORDIMER there are 29 partial paraphrases and 2 global ones, compared to 15 global paraphrases and 13 partial paraphrases in INFOWIN. I believe that this also can be attributed to the

relatively higher degree of freedom a technical translator has in recreating the content of a translation compared to a translator of fiction.

11.4.3 Category shifts

By comparison the operations containing category shifts are found more often in the human translations of computer manuals (ACCESS and EXCEL). Only in one of the texts there are voluntary negations shifts, namely in OS2. Of the recorded simple category shifts (CSHIFT), the most common ones are the following:

1. A subject in the source is translated as an adverbial or prepositional object in the target, for example

(i) SOURCE: **Microsoft Excel** uses …
TARGET: **I Microsoft Excel** används …

(ii) SOURCE: **The line-in jack** can connect the sound adapter.
TARGET: **Med linjeingången** kan du ansluta ljudadaptern.

(iii) SOURCE: **My own heart** must have a feudal compartment.
TARGET: **I mitt eget hjärta** måste det finnas en feodal kammare.

Sometimes these changes are dependent on a different types of verb frame, but for the computer manual translations, most of these examples can be explained by the tendency in Swedish not to personify computer systems (or elements thereof). This is also apparent in the voice changes, which is dealt with later on this chapter. Rather than stating that "system X does this and that for you" the Swedish versions contain many examples of where the translator rephrases this as "With system X you can do this and that". It is a matter of who is in charge, the system or the user, and in Swedish, (judging also from Microsoft's and IBM's own style guides) this way of writing in a "user-oriented" way is advocated explicitly.

2. Adjectival modifiers in the source become prepositional or adverbial modifiers in the target, for example

(iv) SOURCE: The **following** table **…**
TARGET: Tabellen **nedan** …

(v) SOURCE: …translate **medieval** Italian travel narratives …
TARGET: … översätter italienska reseskildringar **från medeltiden**

(vi) SOURCE: The list must contain **labeled** columns.
TARGET: Listan måste innehålla kolumner **med etiketter.**

The category shifts on the clause level (CLSCSHIFT) appear under certain circumstances too. Below the two most common subtypes are listed together with some examples:

1. A temporal adverbial subordinate clause becomes a conditional or causative subordinate clause, for example

(vii) SOURCE: **As** you hide and show …
TARGET: **Om** du visar eller döljer …

(viii) SOURCE: **When** you move…
TARGET: **Om** du flyttar …

(ix) SOURCE: …**when** you hide items …
TARGET: … **om** du döljer element…

2. A subordinate clause functioning as a subject in the source becomes an adverbial in the target, for example

(x) SOURCE: **Replacing a formula**… removes
TARGET: **Om du ersätter en formel**… kommer du att ta bort …

(xi) SOURCE: For example, **putting a top border on the cell below** it, produces …
TARGET: **Om du** exempelvis **placerar en nedre kantlinje på en cell**, ger detta …

It is interesting that the first subtype (of CLSCSHIFT) is so frequent when the possibility to use a straightforward translation is possible. The reason that translators make this choice can only be in terms of speculation and is partly connected to the second subtype above. English constructions such as the ones above are difficult for translators, as there are no obvious grammatical counterparts in Swedish. Phrases such as "replacing a formula" above can indeed be translated by infinitival constructions, such as "att ersätta en formel", but it would not be possible in all contexts. The target examples above could have been rendered as temporal adverbials instead. The possibility that temporal and conditional clauses can be exchanged without any major contextual change in meaning is something that the translator is used to, especially when translating technical documentation where these types of choices have to be made all the time. In other words, in certain contexts temporal and conditional clauses are interchangeable, from the target text point of view, and this is something that translators may transfer as a strategy to other contexts where such segments are translated.

11.4.4 Additions and deletions

What translation segments are added and deleted in translations? From the translation corpus, we can study such operations and see if there are any differences between translation methods used and text types. The data for additions and deletions are shown in Table 81.

Table 81. Expansions and reductions

	EXPADD	EXPDIV	Total expansions	REDDEL	REDCON	Total reductions
Access	23	6	29	42	4	**46**
Excel	20	5	25	35	3	**38**
OS2	49	6	55	47	8	**55**
Infowin	24	2	26	40	6	**46**
Client	5	0	5	26	6	**32**
Gordimer	14	15	**29**	17	3	20
Bellow	6	8	**14**	3	10	13
ATIS	0	0	0	0	3	**3**

All of the operations given in Table 81 entail that the number of translation segments has changed from source to target. For EXPADD (additions) and REDDEL (deletions)it means that a translation segment not present in the source has been added or deleted respectively. For EXPDIV (divergences) and REDCON (convergences), it means that there is a one-to-many or many-to-one relationship between the segments in the source and target text. Usually the presence of an EXPADD or REDDEL operation causes a difference in meaning but not always. If we compare the number of expansions and reductions in the table above, we can note that all computer manual translations have higher values for reductions than for expansions. The translations of the novels show the opposite tendency, namely more expansions than reductions.. Note that expansions and reductions are one major source for a modification of the classification of a translation into a more specific or a less specific sentence translation (cf. section 10.7.2). As was concluded in section 11.3, the manuals have a tendency towards generalisation in the translation (i.e. less specific translations) whereas the novels tend to contain more specific information. From the figures above, it is possible to see that, at least to an extent, these tendencies stems from the number of additions and deletions occurring in the translations. However, in the Gordimer translation, the pattern is not that clear, as there are more deletions than additions. For this particular translation, the major factor for generalisation is to be found among the lexical operations, i.e., the operations that changes the information content for a particular lexical item (see section 11.4.6).

Typical deletions in the manuals are agent deletions and data that the translator possibly judges as implicit in the context. In the example below, the subject/agent "Microsoft Access" is deleted in the translation as well as the information that the instruction has to be performed "in Form view".

(xii) SOURCE: When you paste records into a form **in Form view**, **Microsoft Access** matches the field names of the underlying source table with those of the destination table…
TARGET: När du klistrar in poster i ett formulär passas fältnamnen i den underliggande källtabellen ihop med fältnamnen i måltabellen …

Additions from the fiction translations can be illustrated with a translation from the Bellow novel:

(xiii) SOURCE: To push on to Cairo would have meant the loss of another thousand Israelis and might have caused Russia to intervene.
TARGET: Att rycka fram **ända** till Kairo skulle ha betytt att Israel hade förlorat ytterligare tusen man och kunde ha fått Ryssland att intervenera.

The Swedish adverbial "ända" has no explicit corresponding lexical item in the English source text. The translator has added a piece of extra information (perhaps redundant) to the target text.

11.4.5 Transpositions

Transpositions occur in all the translations except for the ATIS text to a similar degree. In the computer manuals a frequent case of transpositions occurs in the translations of imperative sentences where the imperative verb is not in the initial position, instead this position contains an adverbial. In Swedish, there is a tendency to place the imperative verb in the initial position and therefore the adverbial is often moved from the source position.

(xiv) SOURCE: At its border, **drag** the control that encloses the other controls to a new location.
TARGET: **Dra** i kanten på kontrollen med de inneslutna kontrollerna till kontrollens nya plats.

(xv) SOURCE: To import or attach another table from the same data source, **repeat** Step 6.
TARGET: Upprepa steg 6 om du vill importera eller koppla ytterligare tabeller från samma datakälla.

It seems that the possibility to fill the topic field of an imperative sentence with an adverbial (for example, a conditional clause or a temporal adverbial) is accepted usage, whereas in Swedish, there is a clear tendency to start the sentence with the imperative verb and place the adverbial in another position. Diderichsen (1966, §73, 2nd note) states that the same phenomenon occurs in Danish, namely that the fundament field in imperative sentences is empty in general. Hansen (1987) reports some exceptions to Diderichsen's general claim, but in general, Danish and Swedish seem to be similar with respect to the structure of imperative sentences and the tendency to avoid clauses in front of the imperative verb.

To test whether the phenomenon of the empty fundament field in imperative clauses holds, 581 instances of imperative verbs were extracted from the Stockholm-Umeå Corpus (SUC 1997). Disregarding interjections (e.g. "Nä, Kicki, kom hit" – "No, Kicki, come here") and conjunctions (e.g. "Och var försiktiga – And be careful"), only 12 instances were found where a construction was placed before the imperative verb. Three of them held a short non-clausal adverbial in the initial position, for example:

(xvi) Först och främst – välj ett så enkelt recept som möjligt.
(First and foremost – choose as simple a recipe as possible.)

Two cases were instances of a conditional clause in the initial position followed by the anaphoric "så" (cf. the Danish "så" as described by Hansen (1987)):

(xvii) Om det är du som ringer upp så visa respekt för den andres tid.
(If you are calling, show respect for the other person's time.)

Finally, in seven cases the initial position was filled by a prepositional object ("om X" or "för X"):

(xviii) Om Göta kanal se Gustaf Ekström, Baltzar von Platen and Göta kanal.
(About Göta kanal, see Gustav Ekström, Baltzar von Platen and Göta kanal.)

This last type may seem very specific, and, is to me, a way of writing brief cross-references. I do not want to claim that it is impossible to use adverbial clauses at the beginning of imperative sentences in Swedish, but the tendency is clear, both in the Linköping Translation Corpus and in the SUC corpus: in an overwhelming majority of the cases the initial position is empty in Swedish imperative clauses. This also forms an explanation to some of the transposition operations occurring in the translations as well as to the mood operations.

11.4.6 Lexical operations

The lexical operations together with deletions and additions in general determine to a large extent whether a translation can be judged as more specific, less specific or semantically different translations compared to the source sentence. In Table 82 below the distribution of the different lexical operations in the sample are shown.

Table 82. Distribution of lexical operations

	Access	**XL5**	**OS2**	**InfoWin**	**Client**	**Bellow**	**Gord**	**ATIS**
LexM	12	6	16	9	6	4	18	1
LexL	17	20	51	26	20	1	7	0
LexU	10	13	19	32	9	7	13	9

The tendencies observed in section 11.3 earlier are clear, namely that the computer manuals contain more LEXL operations which contribute to less specific translations and that the two novels contain more LEXM operations which makes the translations more specific. The kind of lexical generalisations present in the computer manuals can be exemplified with the following three sentence pairs (xix-xxi):

(xix) SOURCE: **This process** helps you **identify** the area in the sound graph you want to edit.
TARGET: **Det** hjälper dig att **se** vilka avsnitt som behöver redigeras.

(xx) SOURCE: That way you have access to both **the ID number** and name of the shipper.
TARGET: På så sätt kommer du åt både speditörens **nummer** och namn.

(xxi) SOURCE: In the last dialog box, click **the Finish button**.
TARGET: I den sista dialogrutan klickar du på **Skapa**.

This above example (xxi) is related to the comparatively low WCR and TB values reported for the Microsoft Access text in section 6.3.4. Here we see an illustration of these figures. The classifier "button" is not present in the Swedish translation, probably because the context and the use of the verb ("click") make it redundant to state that the object in question is a "button".

The inverse operation (LEXM), more common in the novels, could be illustrated with the following two examples:

(xxii) SOURCE: Bray offered to be left outside **the Sputnik** in case the other members of the party turned up.
TARGET: Bray föreslog att de skulle sätta av honom utanför **Sputnikbaren** för den händelse några andra deltagare i sällskapet skulle dyka upp.

(xxiii) SOURCE: A fan made a mobile of tiny Viking ships, the sort of **thing** sold in airport shops … bob slowly …
TARGET: En fläkt kom en mobil av små vikingaskepp -- den sortens **souvenirer** som brukar säljas i flygplatsbutiker … att sakta gunga …

Here the translator has made the translation more explicit by making it clear that "the Sputnik" is indeed a bar and that the "things" sold in airport shops are not just any thing, but "souvenirs".

11.4.7 Mood operations

Changes of sentence mood occur in all manuals, but not in the two novels, or in the ATIS text. Only two types of mood shifts are present in the corpus: imperative to declarative (ImpDecl), and declarative to imperative (DeclImp). The distribution is shown in Table 83.

Table 83. Distribution of mood operations

	Access	**XL5**	**OS2**	**InfoWin**	**Client**	**Bellow**	**Gord**	**ATIS**
ImpDecl shifts	20	14	12	9	8	0	0	0
DeclImp shifts	0	0	3	0	1	0	0	0

There is one direction of mood change that dominates, namely the imperative to declarative shift. The examples below can illustrate the typical operation:

(xxiv) SOURCE: If you use a virus protection program on your computer, override it or turn it off before you run the Microsoft Excel Setup program.
TARGET: Om du använder ett program för virusskydd på datorn bör du koppla bort eller stänga av det innan du kör installationsprogrammet för Microsoft Excel.

(xxv) SOURCE: To remove an autoformat at a later time, select a cell within the formatted range, choose the AutoFormat command from the Format menu, and then choose the None option in the Table Format box.
TARGET: Om du vill ta bort ett autoformat vid ett senare tillfälle, markerar du en cell inom det formaterade intervallet, väljer Autoformat på Format-menyn och sedan alternativet "Inget" i rutan "Tabellformat".

(xxvi) SOURCE: To stop drawing shapes, click the button again.
TARGET: När du är färdig klickar du på knappen igen.

In the first of the examples (xxiv) the imperative form is replaced by the insertion of the modal "bör" (Eng. "ought to") which makes the translation declarative in a syntactic sense. The examples xxv and xxvi contain no modal verb insertion in the translation, instead the imperative form is shifted to a declarative Swedish sentence. What is worth noting here is that the adverbial clauses in the initial position is preserved as far as position is concerned, only the sentence moods have changed.

Possibly we see the other side of what was described above in conjunction with transpositions, namely the tendency among the Swedish translators to avoid transferring adverbial clauses in front of imperative verbs and at the same time preserve the imperative sentence mood. Two strategies then seem to crystallise among the translators for handling such source sentences:

1) Keep the sentence mood (imperative) and move the imperative verb to the initial position.

2) Change the sentence mood to declarative and keep the adverbial in the initial position.

The final operation that will be looked at closer is the voice shift.

11.4.8 Voice operations

Sentence voice changes are present in all texts except the ATIS text, see Table 84. In the manuals the majority of the shifts are from active to passive voice, especially notable in the Microsoft manuals.

Table 84. Distribution of voice operations

	Access	**XL5**	**OS2**	**InfoWin**	**Client**	**Bellow**	**Gord**	**ATIS**
ActPass shifts	19	16	12	19	13	3	2	0
PassAct shifts	4	1	11	17	4	2	5	0

Two typical cases of the active to passive shift in the manuals are when (1) the agent (usually a non-animate system or system component) is deleted and (2) the agent is shifted to an adverbial position. These examples are illustrated below:

(xxvii) SOURCE: Microsoft Excel sorts the regions by their March sales amounts.
TARGET: Regionerna sorteras efter försäljningen för mars.

(xxviii) SOURCE: Microsoft Excel uses different terms for some spreadsheet items and actions.
TARGET: I Microsoft Excel används andra termer för vissa kalkylbladselement och funktioner än i andra kalkylprogram.

If the objective is a conscious move to surpress the agent when it represents a system or system component, then shifting the voice to the passive makes it possible for the translator to omit the agent in the translation (see the section on deletions earlier). The other way of avoiding system agents is to shift them to another category and thereby stressing their instrumental function instead, which is illustrated by the second example above. An observation is that the novels do not contain a high number of voice shifts. Instead the voice mood is preserved in most cases, following the style of the author. In the few cases, where there is a shift of voice, it has to do with lexical and stylistic choices and where a voice shift may seem more appropriate for the translator. Often the voice shifts in the fiction material have to do with the choice of particular verb frames, but where the semantic roles underlying the wording are usually preserved ("was given" could be translated to "få" (Eng. "get")) such as in the following example from the Gordimer novel:

(xxix) SOURCE: … he hasn't **been given** a cabinet post.
TARGET: … har han inte **fått** någon plats i regeringen.
(literal transl. he hasn't **received** a post in the cabinet.)

When it comes to voice shifts, the conclusion is that there is a distinct feature present in the manual translations that reveals the translators' objective to avoid placing the system or a system component as an agent. This is achieved by using the passive and either removing the agent altogether or by shifting its role to an instrumental function. Instead of "System X does Y", the translator rephrases it as "Y is done" or as "With system X you can do Y. If the choice was just a lexical one, which it seems to be in the novels, it could be expected that the passive pattern "Y is done by system X" would also be frequent, but it is not.

The inverse operation, changing from the passive to the active, is also present in the corpus, but to a lesser extent than the active to passive shift. Again, it often has to do with whether an agent should be present in the translation or not, but in this case it is the user ("you") that is included in the target text (when this is implicit in the original). Below are two examples where "agentless" source clauses or sentences in the passive have been translated by active sentences and the "user" is made explicit by the subject "du" (Eng. "you").

(xxx) SOURCE: When a field is formatted, the text displayed in the field can differ somewhat from the actual value stored in the table.
TARGET: När **du** angett ett visst format för ett fält kan det hända att värdet visas på ett annat sätt än det sparats i tabellen.

(xxxi) SOURCE: Only one option in a group can be selected at a time.
TARGET: **Du** kan endast markera ett alternativ åt gången i en gruppruta.

11.5 Summary

At the beginning of this chapter five hypotheses were presented. Let us see how these hypotheses hold at this stage:

1. *Translation memories should make translations structurally more equivalent compared to pure human translations.*

This hypothesis cannot be verified or falsified. The two earliest IBM translations were made with the aid of a translation memory tool, but these translations are less consistent structurally than comparable human translations (from Microsoft). The third translation memory translation from IBM is considerably more isomorphic than the first two translations, but it is comparable to the Microsoft translations. The introduction of translation memory software at IBM, Sweden, has probably changed the translation strategy drastically in relative terms. For the IBM translators the change is definitely substantial. The collected and analysed translation data show a clear change for the company, but how far the move towards structural isomorphy will go is still unanswered.

2. *According to Baker (1993) there should be a higher degree of specification in the translations than in the source texts.*

This is shown not be correct, at least not for all text types or for all language directions. The translations of computer manuals are definitely semantically less specific than the translations of fiction. The culture and jargon of the setting for the source text must sometimes be mellowed down. The American style used in computer manuals seems to be a text type that Swedish translators want to "unspecify". Explicit information is taken away in the computer translations, probably in order to fit the instructional culture of the target audience. Baker's claim about explicitation is on the other hand supported by the data from the novel translations, but it does not seem to be a claim that holds generally for all text types.

3. *Automatic translations should have a very high degree of structural and semantic correspondence.*

With a very limited text sample, substantial claims are hard to make, but the hypothesis is clearly validated, as far as the ATIS text goes. As long as MT systems translate sentence-by-sentence, with little or no discourse level analysis or generation, this claim will still be valid. It is very hard to instruct a human or an MT system, to expand or delete information unless the person or MT system understands in the full sense what the text is about.

4. *It should be possible to couple certain translation operations to certain text types.*

By looking at the distribution of specific translation operations, it is clear that changes to sentence mood and voice occur much more frequently in the computer manuals than in the fiction translations. The explanation for these findings is due to the fact that translators of technical texts tend to make the translations more user-centred and to a certain extent take away some of the focus on the computer system or system components. Such an overall trend is

then manifested linguistically by changing sentence voice to the passive and surpressing the agent if the subject of the active sentence holds a "system agent". Furthermore, when a shift from the passive to the active occurs, it is never a system component that is made the subject in the target sentence, instead it is the user ("you") that is put in focus.

Another type of translation operation that only occurs in the computer manual translations is the mood shift. In Swedish writers tend to be biased to putting the imperative verb in the initial position of a sentence. When imperative sentences in English begin with something other than the verb (for example, adverbials), the translators tend to adopt two different strategies: (1) move the imperative verb to the initial position and move the adverbial to a later position in the target sentence; and (2) keep the adverbial in the initial position and change the target sentence into declarative mood. In lists of instructions, this could mean that symmetry in style (in the case of imperative sentences) is sacrificed for readability, namely when the translator chooses to translate one or more of the imperative instructions into declaratives. But usually this takes place when it is important that the reader is first confronted with a fronted adverbial, like a temporal or conditional clause. The mood shifts consequently often co-occur with transpositions.

5. *The discovery of the apparently ineffecient use of translation memory-based translation tools, uncovered with the analysis tools discussed in chapter 6, should also be visible in a closer structural and semantic correspondence.*

The characterisation of the OS2 and INFOWIN translations as more communicative and freer translations is made even clearer by the correspondence analysis. Compared to the later translation of the CLIENT text, both the OS2 and INFOWIN translations show considerably more traces of both structural and semantic changes. In other words, when the IBM translators have adopted (or become used to) the translation memory tools, the translation of the IBM manuals tend to become more source-oriented than in the two first translations.

A question for the correspondence model is to what extent it could be applied in a more automatic fashion in the future. The most time-consuming and difficult part of its application is the actual tagging work, which in this study was done manually. Here it is possible that the tags describing structure and syntactic function could be arrived at automatically, at least partially, with the aid of syntactic taggers like the Constraint Grammar taggers (Karlsson et al. 1995, Tapanainen 1996). However, to be able to decide semantic relationships as far as degrees of specification and decisions about paraphrases are concerned, it will probably be difficult to arrive at such information without using the knowledge from human sources.

12 Summary and discussion

Parallel corpora and their applications have been investigated in this thesis. In particular I have shown different ways in which translation corpora can be taken as the empirical foundation for improving the understanding of translation in respect to relationships between source and target texts. The improved understanding of what these relationships are would then be put to use in translation systems and translation tools.

The explicit empirical foundation for this work is the Linköping Translation Corpus (LTC), which consists of eight English-Swedish translations (altogether just over 1,500,000 words). LTC contains translations of computer software manuals, made both with the aid of translation memory software as well as without any computer aids. The corpus also contains two novels (translated in the traditional way) and one short translation of dialogue which was produced by an automatic translation system.

In order to build and examine the translation corpus a number of tools have been developed:

- Tools for aligning paragraphs and sentences in order to build translation corpora
- Diagnostic tools to analyse recurrence profiles of source and target texts
- Discrepancy analysis to analyse translation consistency
- Extraction of multi-word units (terminology and phrases) (Frasse-1 and Frasse-2)
- Bilingual concordancing
- Bilingual word alignment for full text alignment and bilingual lexicon extraction (LWA)
- Tools for the evaluation of word alignment systems (PLUG Link Annotator and Link Scorer).

The above tools are all *knowledge-lite*, in the sense that they do not rely on large and complex linguistic resources, such as lexicons, parts-of-speech information and grammars. A knowledge-lite approach is not an end in itself, but it has the advantages of making systems more portable to other languages and text types. Only a minimal amount of time has to be spent on specifying language- or text-specific knowledge. The disadvantages of knowledge-lite techniques are slightly lower performance and a loss of power of expression.

The main contribution of the tools listed above is the combined functionality that they provide. Other tools are available that can do the same things as some

of these tools (for example, regarding sentence alignment and bilingual concordancing). However, the discrepancy tool is in itself an original contribution as well as the knowledge-lite approach for LWA and the combination of language filters and entropy thresholds in Frasse-2. Furthermore, the flexibility and possibility of extended applications of the PLUG Link Annotator are new contributions to the field.

To be able to describe more complex types of relationships between units in a source text and a target text, a more elaborate model is required. In chapter 10 and 11, a model for describing structural and semantic correspondences in the Linköping Translation Corpus was described and applied. The data used and the analyses were made by hand as currently there are no available techniques for supplying translations with the required knowledge.

Throughout the thesis a number of *measures* have been formulated and applied in order to describe characteristics present in either source or target texts and relationships between units in source and target texts. These measures include recurrence rates in monolingual texts, co-occurrence ratios, measures for translation consistency as well as measures for structural and semantic changes between a source text and its target text.

12.1 Translation characteristics related to text type

A comparison of *source texts* for software manuals with source texts for fiction shows that the manuals contain more repetitions, more multi-word units (expressed as technical terms) and a more restricted vocabulary (low word type/token ratios). The data on recurrence rates and number of useful MWUs were compiled with the aid of components in the Dave toolbox and with Frasse-2. Data on sentence mappings can be one source of indicating how source-oriented a translation is. If one source sentence is translated by exactly one target sentence the sentence mapping is said to be 1-1 and the number of 1-1 mappings can give a relative indication of the extent of source-orientation in a translation. The translations in the LTC contain 1-1 sentence mappings in the region of 96.36 to 98.35 per cent, with the exception of one of the IBM texts that had a 87.53 per cent mapping rate (OS/2 User's Guide). This difference is in itself an indication that the IBM text is not as source-oriented as the other texts. With the aid of the bilingual concordancing tool, measures for word co-occurrence and translation bias were calculated, and the tendency that the OS/2 translation was the least source-oriented translation was confirmed by having the lowest scores for word co-occurrence and translation bias.

By using measures for structural and semantic changes from source to target, several interesting text type differences were discovered. The structural changes were on average smaller in the fiction translations compared to the translations of computer manuals. Another notable difference is that the translations of the computer manuals are definitely semantically less specific than the translations of the novels. Explicit information is made more general or removed in the Swedish translations of the computer manuals. The culture and jargon of the setting of a source text must be revised in order to fit the target culture, for

example when it comes to instructional passages or text sections that are connected to marketing. The claim by Baker (1993) about "explicitation", that is, that there should be a higher degree of specification or explicitness in the translations compared to source texts, is on the other hand supported by the data from the novel translations. In the novels, the translators have, in general, added information in the translations. However, the analysis of the computer manual translations show that "explicitation" is *not* a universal feature in translation.

Certain *translation operations* can also be tied to certain text types. Changes to *sentence mood* and *voice* occur much more frequently in the computer manuals than in the fiction translations. The explanation for these findings is that the translators of these technical texts tend to make the translations more user-centred and present a different view of the computer system and system components. This general trend is then manifested linguistically by changing the sentence voice to the passive and surpressing the agent if the subject of the active sentence holds a "system agent" in the source sentence. Furthermore, when a shift from the passive to the active occurs, it is never a system component that is made the subject in the target sentence, instead it is the user ("you") that is put in focus.

Another type of translation operation that only occurs in the manual translations is the mood shift. In Swedish, writers tend to be biased to put the imperative verb in the initial position of a sentence. When imperative sentences in English begin with something other than the verb (for example, adverbials), the translators tend to adopt two different strategies: (1) move the imperative verb to the initial position and move the adverbial to a later position in the target sentence; and (2) keep the adverbial in the initial position and change the target sentence into declarative mood. In lists of instructions, this could mean that symmetry in style is sacrificed for readability, i.e., when the translator chooses to translate one or more of the imperative instructions into declaratives. Usually this takes place when it is important that the reader is first confronted with a fronted adverbial, like a conditional or temporal clause. The mood shifts consequently often co-occur with transpositions.

12.2 Translation characteristics related to translation methods

One of the prevailing techniques in the translation industry are translation memory-based tools, which are based on the notion of *reusability*. Every time a translation unit (be it a sentence or a phrase) reappears in a source text, the translator should be able to reuse previously made translations and thereby save time and ensure the use of correct terminology and phraseology. For translations made with translation memory-based systems, the hypothesis was that this translation method would result in more source-oriented translations compared to translations made without such tools. The hypothesis was based on the assumption that a target-oriented approach would lead to decreased reusability, as the recycled segments should be used in different contexts. But the hypothesis for increased source-orientation using translation memories could not be verified. The first IBM translations made with translation memories actually exhibit more of a target-orientation than any of the other texts in LTC.

However, over time the IBM translations show a tendency of a higher degree of source-orientation, but there are no apparent differences as far as structure and content are concerned compared to the translations of the Microsoft texts (which were done without any translation tools). The discrepancy analyses of the translations strengthened this observation even further. The first attempts at using translation memories at IBM did not live up to the expectations as no efficiency gains can be observed in these translations compared to the human translations of the Microsoft texts. In interviews with the IBM translators, this observation was confirmed and was explained by a clash of new technology and the traditional translation culture at the company. In the eighties, the IBM translators more or less rewrote the American manuals in Swedish, making the translations extremely target-oriented, or more or less Swedish adaptations of American originals. The introduction of translation memory tools at the company then clashed with the existing translation culture, which resulted in openly negative attitudes towards the new technology and caused some translators to stick to the old way of translating. Initially there were also technical and administrative problems, for example with the distribution and sharing of translation memories. This also contributed to the resistance from the translation teams that were supposed to use the tools.

Only one of the translations in LTC had been done by an automatic translation system (the ATIS text). This text has been shown to be more or less 100 per cent source-oriented, with no apparent changes to structure and content. The translation can be said to find the shortest possible way from source to target, with only the necessary adjustments to make the target sentence grammatically correct. With only one short sample from automatic translation it is difficult to claim any general results. But as long as MT systems translate sentence-by-sentence, on a rule-for-rule basis, with little or no discourse-level analysis or generation, automatic translations will by default be very similar in structure and content to the source text.

12.3 Translation memories and reusability

In chapter 3, it was showed that the basic principle of reusability in translation is applicable, i.e., that most recurrent sentences can be recycled, according to translators. It is only when sentences occur in different functional contexts that this principle has to be abandoned, for example, if the same segment occurs as a header or in a table cell. This is good news for the users of translation memory tools, but the bad news is that the survey also showed that translators have difficulties in accepting suggested translations. Alternatively, one can say that translators do not share a common view of what is to be regarded as the "best" translation alternative if there are several candidates to choose from. Consequently, it is likely that translators will want to change already stored translations, which will decrease efficiency and make translations less consistent if several translators work with the same source text. This observation is one explanation of what happened when translation memory tools were introduced at IBM. Translators have difficulties in accepting previous translations as they are.

In order to overcome the difficulties that translators have with reusing standard translations, more emphasis should be put on the verification and administration of translation memories. With improved functions for cleaning-up translation memories before they are archived, it would be possible to give the translators a greater degree of freedom to choose the optimal translation, which in turn would cause a more positive attitude towards the technology. At the post-editing stage, one standard translation would be archived, and erroneous variant translations would be removed from the archive.

12.4 Automatic translation

Translation corpora will improve machine translation systems in several ways. For rule-based systems it is possible to incorporate translation operations such as the ones found in the computer manual translations in LTC. Shifts regarding sentence mood and sentence voice could therefore be handled by MT systems.

Automatic extraction of bilingual lexicons for MT systems can already be made, and is a pre-requisite for swift porting of systems to new domains. In order to get closer to a more target-oriented translation, further and more detailed studies on correspondences would have to be carried out. In this thesis, certain operations connected to voice and mood shifts as well as deletions and additions have been identified for the translations of computer manuals. But the general trend, that English originals tend to loose explicit information in the Swedish translations, requires a more elaborate understanding of what is taking place on the discourse level and what adaptation to a genre- and text-type specific norm really means.

12.5 Multi-word unit extraction

Multi-word unit (MWU) extraction for monolingual texts is important in the area of lexicography and terminology, and also as a resource for different types of systems, such as word alignment systems and machine translation systems. In this thesis, two variants of a knowledge-lite MWU extraction program have been presented: Frasse-1 and Frasse-2. They are knowledge-lite in the sense that no abstract linguistic information is required. Both programs operate on the text seen as strings of tokens and are combined with language filters. Frasse-1 uses only a frequency threshold and in Frasse-2 an entropy threshold and a more elaborate language filter are adopted. The systems are fast because of the knowledge-lite approach and easy to apply to new languages and new texts because the only resource that needs adjusting is the language filter. Frasse-1 will produce more MWUs than Frasse-2, but with a lower precision. Frasse-2 is therefore more useful when no human input or revision of the MWU resources can be made. When human revision of the MWU lists is possible, the output from Frasse-1 (or a combination of the output from Frasse-1 and Frasse-2) can be filtered manually to arrive at a larger number of accurate MWUs. The Frasse systems have been successfully applied to Swedish, English and German texts.

12.6 Word alignment and evaluations

The Linköping Word Aligner (LWA) is a knowledge-lite word alignment system that produces full-text alignment and extracts a bilingual lexicon. The system is built around a statistical engine where a word association score is used to find correspondences between single or multi word units from a source and a target text. The input to the system is a sentence-aligned translation corpus (bitext). The statistical engine in the system is complemented with four modules that handles MWUs, morphology, closed-class/open-class expressions and position weights. Several filters and tests are also possible to include or exclude in the linking process, for example strategies for linking unique words and cognates, etc.

To be able to evaluate LWA (and other word alignment systems), an annotation program, called the PLUG Link Annotator, was developed. One or several annotators can use the PLUG Link Annotator to create reference data (gold standards) against which the output of a word alignment system can be evaluated. The selection of which source words that are used in the annotation process can be made in different ways, for example, by random sampling of text tokens, frequency-balanced word-sampling or by sampling only content words. A separate module, the Link Scorer, compares the reference data with the output from the word alignment program and delivers a set of scores for the performance of the system.

LWA has been evaluated in three ways. The first evaluation was made by evaluating a sample of the extracted lexicon from some of the translations in the Linköping Translation Corpus. The second evaluation was made in conjunction with an international word alignment test, called ARCADE (Véronis and Langlais 1999), where LWA was applied to French/English translations. In the ARCADE project, the evaluation was made by using prior reference material, a gold standard, and all participating systems were compared for recall and precision scores. In the third evaluation, different system configurations of LWA were tested with the use of several types of gold standards. The gold standards were created with the aid of the PLUG Link Annotator, and the scores for the different configurations were automatically calculated by comparing the different outputs with the gold standards. The conclusions from the evaluations are that LWA stand up well in the Arcade project, in spite of the knowledge-lite approach, and that the swift porting of the system to a new language pair was indeed made possible precisely by the fact that LWA is knowledge-lite. It was further concluded that precision and recall scores for different systems and system configurations can only be compared if the same bitext is used as input and furthermore that the gold standard used to evaluate the output is created with comparable criteria. Using different sampling methods for the source items in the gold standard will influence the precision and recall scores. For example, random sampling of source tokens for the gold standard will produce higher recall results and lower precision results compared to a frequency-balanced sampling of content words only.

All evaluations show that a purely statistical approach can be improved by adding different knowledge-lite modules. For example, the number of type links

increases by more than 300 per cent when all modules are activated compared to the statistical baseline configuration. The accuracy of the extracted links is also improved with the modules.

12.7 Lexicography and contrastive linguistics

For lexicography and contrastive linguistics, there is vast amount of knowledge hidden in parallel corpora. By using bilingual concordance programs, it is possible to study how source units are related to target units in various ways. Depending on the task at hand, both the lexicographer and the contrastive linguist can study the distribution of lexical items and more complex constructions for applied or research purposes. One difference between a bilingual concordancer and a word alignment system is that for a bilingual concordancer the user is investigating something specified, be it a source word or a combination of a source or a target word. For a word alignment system, the search is not specified; the system will try to output anything that is deemed as a probable pair of source unit and target unit. Then it is up to the user to decide whether the relationship is usable or interesting. It is exactly this characteristic that makes word alignment system such an interesting tool for lexicographers. Given large translation corpora and a word alignment system, tens of thousands of potential entries for bilingual lexicons can be extracted automatically. The lexicographer can then compare the automatically extracted lexical pairs with an existing bilingual lexicon and decide what and what not to include in a future release of the lexicon.

12.8 Translation studies

From the field of translation studies, the interest in translation corpora has been increasing over a number of years. The emergence of powerful hardware, both for storage and processing, has made it possible to collect and investigate much larger amounts of texts than ever before. However, with the developments within natural language processing and corpus linguistics (sentence alignment, parts-of-speech tagging, word alignment, concordancers, etc.) the ground has been paved for bringing the empirical foundation as it is manifested in real translations onto the translation scholars desk. Hypotheses concerning for example, universals in translations can now be tested on a much wider scale.

Various interesting areas within translation studies could be approached given the appropriate corpora. Patterns of "spoken" language could be investigated by looking into the dialogue of stage plays. The characteristics underlying screen and TV translation could be made clearer by having translation corpora with such material.

The results that translation scholars produce need not, however, remain within the translation studies field. It is of vital importance that both general and more genre- or text-type specific knowledge also are brought to the attention of developers of machine translation systems and translation tools.

12.9 Future work

Below I list some of the areas or applications that I regard as important continuations of this work:

- *Improved translation memory tools.* Modules for terminology extraction and word alignment can be integrated with the TM tools. Validation tools for consistent terminology translations and compliance to existing style guides would also be a considerable improvement.

- *Improved automatic translation systems.* Rule-based MT systems can be improved by integrating empirically validated translation shifts into transfer modules. Another area is to continue the path on hybrid systems where example-based and rule-based MT approaches are combined in order to improve the quality of the target texts.

- Testing the *differences between knowledge-lite and knowledge-intensive translation processing.* What can be achieved when knowledge-lite approaches are used and what kind of processing requires more detailed linguistic resources? This is important, for instance, for the extraction of terms and collocations.

- *Lexicography and parallel texts.* Develop techniques and methods for using parallel text corpora in work with lexicon construction.

- *An integration of bilingual concordancing and lexicons* where the output from word alignment systems are taken as the input to the bilingual concordancer. The user should then be able to browse and search the translations from different perspectives. For example, by entering a bilingual lexicon, the user could click on a lexical entry and view all the instances in the translation corpus where this particular link appears. Furthermore, when the user specifies a search for a particular source word or expression, the concordance component will also highlight the corresponding translation in the target text (if the system has found one).

- *Translation databases.* Parallel texts containing more explicit translation knowledge, such as correspondences, syntactic and semantic information, etc. This can be seen as an automatic extension of the correspondence model described in chapter 10.

Bibliography

Primary text sources

English originals

IBM InfoWindow II 3488 User's Guide, IBM Corporation, 1993.

IBM OS/2 2.1 Using the Operating System, IBM Corporation, 1993

IBM OS/2 2.1 Installation Guide, IBM Corporation, 1993

IBM Client Access for Windows 3.1 User's Guide, IBM Corporation, 1995

Microsoft Excel for Windows 5.0 User's Guide, Microsoft Corporation, 1993

Microsoft Access for Windows 2.0 User's Guide, Microsoft Corporation, 1994.

Bellow, Saul, *To Jerusalem and back: a personal account*, Viking P. for the Jewish Publication Society of America, New York, 1976.

Gordimer, Nadine, *A Guest of Honour*, Cape, London, 1971.

Swedish translations

IBM InfoWindow II 3488 Användarhandbok, IBM Corporation, 1993.

IBM OS/2 2.1 Användarhandbok, IBM Corporation, 1993

IBM OS/2 2.1 Installationsanvisningar, IBM Corporation, 1993

IBM Client Access för Windows Användarhandbok, IBM Corporation, 1995

Microsoft Excel för Windows 5.0 Användarhandbok, Microsoft Corporation, 1993

Microsoft Access för Windows 2.0 Användarhandbok, Microsoft Corporation, 1994.

Bellow, Saul, *Jerusalem tur och retur*. Bonniers, översättning Caj Lundgren, Stockholm, 1977.

Gordimer, Nadine, *Hedersgästen*, Bonniers, översättning Magnus K:son Lindberg, Stockholm 1977.

Agnäs, M.-S., H. Alshawi, D. Carter, K. Ceder, M. Collins, R. Crouch, V. Digalakis, B. Ekholm, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, P. Price, S. Pulman, M. Rayner, C. Samuelsson and T. Svensson (1994). Spoken Language Translator: First-Year Report. Stockholm, Swedish Institute of Computer Science.

Adriaens, G. and L. Macken (1995). Technological evaluation of a controlled language application: precision, recall and convergence tests for SECC. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation.* Leuven: 123-141.

Ahrenberg, L. and M. Merkel (1996). On Translation Corpora and Translation Support Tools: A Project Report. *Languages in Contrast. Text-based cross-linguistic studies.* Aijmer, K., B. Altenberg and M. Johansson (eds.). Lund, Lund University Press: 185-198.

Ahrenberg, L. and M. Merkel (1997). Språkliga effekter av översättningssystem. *Svenskan i IT-samhället.* O. Josephson (ed.). Uppsala, Hallgren & Fallgren: 96-116.

Ahrenberg, L., M. Andersson and M. Merkel (1998a). A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98.* Montreal: 29-35.

Ahrenberg, L., M. Andersson and M. Merkel (1999). A knowledge-lite approach to word alignment. To be published in J. Véronis (ed.), *Parallel Text Processing*, Kluwer Academic Press.

Ahrenberg, L., M. Merkel, D. Ridings, A. Sågvall Hein and J. Tiedemann. (1998b). Automatic processing of parallel corpora: A Swedish perspective. Plug Report. Also published in Linköping University Electronic Press, Computer and Information Science Series 1999/002.

Aijmer, K. (1998). Epistemic predicates in contrast. *Corpora and cross-linguistic research.* S. Johansson and S. Oksefjell. Amsterdam, Rodopi: 277-295.

Aijmer, K. (1999). Epistemic Possibility in an English-Swedish Perspective. *Out of corpora. Studies in honour of Stig Johansson.* H. Hasselgård and S. Oksefjell. Amsterdam, Rodopi: 301-326.

Altenberg, B. (1998). Connectors and sentence openings in English and Swedish. *Corpora and cross-linguistic research.* S. Johansson and S. Oksefjell. Amsterdam, Rodopi: 115-143.

Altenberg, B. (1999). Adverbial connectors in English and Swedish. *Out of corpora. Studies in honour of Stig Johansson.* H. Hasselgård and S. Oksefjell. Amsterdam, Rodopi: 249-268.

Apple (1994). *Apples terminologihandbok.* Stockholm, Apple Computer AB.

Arnold, D., L. Balkan, R.L. Humphries, S. Meijer and L. Sadler. (1994). *Machine Translation - An Introductory Guide*. Oxford, Ncc Blackwell Ltd.

Atkins, S. and J. Clear (1992). "Corpus Design Criteria." *Literary and Linguistic Computing* 7(1): 1-16.

Barlow, M. (1995). "Paraconc: A Parallel Concordancer for Parallel Texts". *Computers and Texts 10*: 14-16.

Baker, M. (1993). Corpus Linguistics and Translation Studies – Implications and Applications. *Text and Technology*. M. Baker, G. Francis and E. Tognine-Bonelli (eds.). Philadelphia/Amsterdam, John Benjamins Publishing Company: 233-252.

Baker, M. (1995). "Corpora in Translation Studies - An Overview and Some Suggestions for Future Research." *Target* 7((2)): 223-243.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Terminology, LSP and Translation*. H. Somers (ed.). Amsterdam, Benjamins: 175-186.

Bowker, L., M. Cronin, D. Kenny and J. Pearson (1998). *Unity in Diversity. Current trends in translation studies*. Manchester, St. Jerome Publishing.

British National Corpus (1999). *The British National Corpus*, http://info.ox.ac.uk/bnc/index.html (Access date: March 16, 1999).

Brown, P.F., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer and P. Roosin (1988). A Statistical Approach to Language Translation. *Proceedings of the 12th International Conference on Computational Linguistics*. Budapest: 71-76.

Brown, P.F., J.C. Lai, and R.L. Mercer (1991). Aligning sentences in parallel corpora. *Proceedings from 29th Annual Meeting of the Association of Computational Linguistics (ACL-91)*: 169-176.

Brown, R.D. (1996). Example-Based Machine Translation in the Pangloss System. *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen: 169-174.

Catford, J.C. (1965) *A Linguistic Theory of Translation Oxford*: Oxford University Press.

Chandioux, J. and A. Grimaila (1996). "Specialized" Machine Translation. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*: 206-211.

Chang, J.J.S. and S.-J. Ker (1996). Aligning More Words with High Precision for Small Bilingual Corpora. *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen: 210-215.

Chen, S.F. (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information. *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*. Columbus, Ohio: 9-16.

Chen, K.-H. and H.-H. Chen (1994). Extracting Noun Phrases from Large-Scale Texts: a Hybrid Approach and Its Automatic Evaluation. *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics*. New Mexico: 234-241.

Choueka, Y. (1988). Looking for needles in a haystack. *Proceedings from RIAO 88, User-oriented Content-based Text and Image Handling*: 609-623.

Church, K.W. and W.A. Gale (1995). Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge: 121-130.

Church, K.W. (1993). A Program for Aligning Parallel Texts at the Character Level. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*: 1-8.

Church, K.W. and P. Hanks (1990). "Word association norms, mutual information and lexicography." *Computational Linguistics* 16(1): 22-29.

Church, K.W. and E.H. Hovy (1993). "Good Applications for Crummy Machine Translation." *Machine Translation* 8(1): 239-258.

Dagan, I. and K.W. Church (1994). Termight: Identifying and Translating Technical Terminology. *Proceedings from the Conference on Applied Natural Language Processing (ANLP-94)*: 34-40.

Dagan, I. and K.W. Church (1997). "Termight: Coordinating Humans and Machines in Bilingual Terminology Translation." *Machine Translation* 12(1-2): 89-107.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *Proceedings of the workshop The Balancing Act - Combining Symbolic and Statistical Approaches to Language*:29-36.

Danielsson, P. and K. Mühlenbock (1998). When Stålhandske becomes Steelglove: A Corpus Based Study of Names in Parallel Text. *Machine Translation and the Information Soup*. D. Farwell, L. Gerber and E. Hovy. Berlin, Springer: 266-274.

Dechert, H. W. and U. Sandrock (1986). Thinking-Aloud Protocols: The Decomposition of Language Processing. V. Cook (ed.), *Experimental Approaches to Second Language Learning*. Oxford: Pergamon:111-126.

Diderichsen, P. (1966). *Elementær dansk Grammatik*. Copenhagen, Gyldendal.

Flanagan, M. (1996). Two Years Online: Experiences, Challenges, and Trends. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 192-197.

Flanagan, M. (1997). "MT Today: Emerging Roles, New Successes." *Machine Translation* 12(1-2): 25-27.

Fluhr, C., F. Bisson and F. Elkateb (1999). Mutual benefit of sentence/word alignment and crosslinugal information retrieval. To be published in *Parallel Text Processing*. J. Véronis, Kluwer Academic Press.

Foster, G., P. Isabelle and P. Plamondon (1997). "Target-Text Mediated Interactive Machine Translation". *Machine Translation* 12(1-2):175-194.

Francis, W.N. and H. Kucera (1964). Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers. Providence R.I., Department of Linguistics, Brown University.

Fung, P. (1995). Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge: 173-183.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-95)*. Cambridge, Massachusetts: 236-243.

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. *Machine Translation and the Information Soup*. D. Farwell, L. Gerber and E. Hovy. Berlin, Springer: 1-17.

Fung, P. and K.W. Church (1994). K-vec: A New Approach for Aligning Parallel Texts. *Proceedings from the 15th International Conference on Computational Linguistics (Coling-94)*. Kyoto: 1096-1102.

Fung, P. and K. McKeown (1997). "A Technical Word and Term Translation Aid using Noisy Parallel Corpora across Language Groups." *Machine Translation* 12(1-2): 53-87.

Fung, P. and D. Wu (1994). Statistical Augmentation of a Chinese Machine-Readable Dictionary. *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*. Kyoto: 69-86.

Gale, W. and K.W. Church (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*: 177-184.

Gaussier, E. (1998). Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora. *Proceedings of COLING-ACL '98*. Montreal: 444-450.

Gdaniec, C. (1994). The Logos Translatability Index. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, AMTA: 97-105.

Gdaniec, C. (1998). Lexical Choice and Syntactic Generation in a Transfer System: Transformations in the New LMT English-German System. *Machine Translation and the Information Soup*. D. Farwell, L. Gerber and E. Hovy. Berlin, Springer: 408-420.

Gdaniec, C. and P. Schmid (1995). Constituent Shifts in the Logos English-German System. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*. Leuven: 311-318.

Gellerstam, M. (1985). Translationese in Swedish novels translated from English. *Proceedings of the Scandinavian Symposium on Translation Theory (SSOTT) II*. L. Wollin and L. H, Liber förlag, Malmö: 88-95.

Gellerstam, M. (1993). "Nej", sade hon iskallt - Dialoger i ett tvärvetenskapligt perspektiv. *Mänsklig mångfald*, Göteborgs universitet. 6: 113-126.

Gellerstam, M. (1996). Translation as a source for cross-linguistic studies. *Languages in Contrast*. K. Aijmer, B. Altenberg and M. Johansson, Lund University Press: 53-62.

Granlund, F. (1999*). Utvärdering av maskinöversättning med Engelska hjälpredan*. Manuscript. Institutionen för lingvistik, Uppsala universitet.

Grishman, R. (1994). Iterative Alignment of Syntactic Structures for a Bilingual Corpus. *Proceedings of the Second Annual Workshop for Very Large Corpora*. Kyoto: 57-68.

Haaland, L. (1997). *Contrastive Linking in English and Norwegian*. Unpublished MA thesis, Department of British and American Studies, University of Oslo.

Haas, W. (1968). The Theory of Translation. *The Theory of Meaning*. G. H. R. Parkinson. Oxford, Oxford University Press: 86-108.

Hann, M. (1992). *The Key to Technical Translation, Volume One - Concept Specification*. Amsterdam, John Benjamins Publishing Company.

Hansen, E. (1987). Imperativens fundamentfelt - Et råmateriale. *Sætningsshemæt og dets stilling - 50 år efter*. L. Heltoft and J. E. Anderson, Akademisk Forlag: 99-104.

Harris, B. (1988). "Bi-text, a new concept in translation theory." *Language Monthly* 54: 8-10.

Hartmann, R.R.K. (1997). From Contrastive Textology to Parallel Text Corpora: Theory and Applications. *Language History and Linguistic Modelling A Festschrift for Jacek Fisiak*, R. Hickly and S. Puppel (eds.), Berlin: M. de Gruyter: 1973-1987.

Hasselgård, H. (1996). Some methodological issues in a contrastive study of word order in English and Norwegian. *Languages in Contrast. Textbased cross-linguistic studies.* Eds. Aijmer, K., Altenberg, B. Johansson, M. Lund, Lund University Press: 113-126.

Hasselgård, H. and S. Oksefjell (1999). *Out of corpora. Studies in honour of Stig Johansson.* Amsterdam, Rodopi.

Heyn, M. (1996). "Present and Future Needs in the CAT World." *The Localisation Industry Standards Association Forum Newsletter* 5(3): 15-33.

Heyn, M. (1998). Translation Memories: Insights and prospects. *Unity in Diversity. Current trends in translation studies.* Eds. L. Bowker, M. Cronin, D. Kenny and J. Pearson. Manchester, St. Jerome Publishing: 123-136.

Holm, M. and M. Olsson (1996). En översättningsredigerares arbetsbänk. Effekter av och erfarenheter från integration och vidareutveckling av verktyg för datorstödd översättning. Master's Thesis. Department of Computer and Information Science, Linköping University.

Holmes, J.S. (1972). *The Name and Nature of Translation S*tudies. Amsterdam: Translation Studies Section., Univeristy of Amsterdam, Department of General Literary Studies. (reprinted Holmes (1988:81-91).

Holmes, J.S. (1988). *Translated!: Papers on Literary Translation and Translation Studies.* Amsterdam: Rodopi.

Hull, D.A. (1998). A Practical Approach to Terminology Alignment. *Proceedings of Computerm '98 (First Workshop on Computational Terminology).* Montreal: 1-7.

Hunt, J.W. and T.G. Szymanski (1977). "A Fast Algorithm for Computing Longest Common Subsequences." *Communications of the ACM* 20(5): 350-353.

Hutchins, W.J. (1986). *Machine Translation: Past, Present, Future.* Chichester, Ellis Horwood Limited.

Hutchins, W..J. (1996). The State of Machine Translation in Europe. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96).* Montreal: 198-205.

Ingo, R. (1991). *Från källspråk till målspråk - Introduktion i översättningsvetenskap.* Lund, Studentlitteratur.

Isabelle, P., M. Dymetman G. Foster, J.-M. Jutrac, E. Macklovitch, F. Perraul, X. Ren and M. Simard (1993). Translation Analysis and Translation Automation. *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93).* Kyoto: 201-217.

Johansson, C. (1996). Good Bigrams. *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96).* Copenhagen: 592-597.

Johansson, S. and K. Hofland (1994). Towards an English-Norwegian Parallel Corpus. *Creating and Using English Language Corpora.* U. Fries, G. Tottie and P. Scheider. Zürich, Editions Rodopi: 25-37.

Johansson, S. and S. Oksefjell (1998). *Corpora and cross-linguistic research.* Amsterdam, Rodopi.

Jones, D. (1992). Non-hybrid example-based machine translation architectures. Proceedings of TMI-92. Montreal:163-71.

Jones, D. and M. Alexa (1997). Toward automatically aligning German compounds with English word groups. *New Methods in Language Processing.* D. Jones and H. Somers. London, UCL Press: 207-218.

Kaji, H. and T. Aizone (1996). Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96).* Copenhagen: 23-28.

Karlgren, J. and D. Cutting (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *Proceedings from the 15th International Conference on Computational Linguistics (Coling-94).* Kyoto: 1071-1075.

Karlsson, F. A. Voutilainen, J. Hekkilä and A. Anttila (1995). *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* Berlin and New York: Mouton de Gruyter.

Kay, M. (1980). *The Proper Place of Men and Machine in Language Translation.* , CSL-80-11, Palo AltoXerox. Reprinted in *Machine Translation* 12(1-2): 3-23.

Kay, M. and M. Röscheisen (1993). "Text-Translation Alignment." *Computational Linguistics* 19(1): 121-142.

Kita, K., T. Omoto Y. Yano and Y. Kato. (1994). Application of Corpora in Second Language Learning - The Problem of Collocational Knowledge Acquisition. *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2).* Kyoto: 43-56.

Kitamura, M. and Y. Matsumoto (1996). Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. *Proceedings of the Fourth Workshop on Very Large Corporaa.* E. Ejerhed and I. Dagan. Copenhagen: 79-87.

Klavans, J. and E. Tzoukermann (1990). The BICORD System. Combining Lexical Information from Bilingual Corpora and Machine Readable Dictionaries. *Proceedings of the 13th International Conference of Computational Linguistics (Coling-90).* Helsinki: 174-179.

Kupiec, J. (1993). An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics (ACL-93)*: 17-22.

Langlois, L. (1996). A New Tool for Bilingual Lexicographers. *Proceedings from the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 34-42.

Larsson, A. and M. Merkel (1994). Semiotics at Work: Technical Translation and Communication in a Multilingual Corporate Environment. *Papers from Nordiska Datalingvistikdagarna (NODALIDA)*. Stockholm: 155-164.

Laviosa, S. (1998). The English Comparable Corpus: A Resource and a Methodology. *Unity in Diversity - Current Trends in Translation Studies*. L. Bowker, M. Cronin, D. Kenny and J. Pearson. Manchester, St. Jerome Publishing: 101-112.

Leech, G. (1997). Introducing Corpus Annotation. *Corpus Annotation - Linguistic Information from Computer Text Corpora*. R. Garside, G. Leech and A. McEnery. London, Addison Wesley Longman: 1-18.

Lindquist, H. (1989). *English Adverbials in Translation - A Corpus Study of Swedish Renderings*. Lund, Lund University Press.

Lörcher, W. (1992). "Investigating the translation process." *Meta* 37(3): 426-439.

Macklovitch, E. (1992). Corpus-based Tools for Translators. *Proceedings of the 33rd Annual Conference of the American Translators Association*, San Diego: 317-328.

Macklovitch, E. (1994). Using Bi-textual Alignment for Translation Validation: the TransCheck System.*Proceedings of the First Conference of the Association for Machine Translation in the Americas AMTA-94*: 157-168.

Macklovitch, E. (1995). *Can Terminology Consistency be Validated Automatically?* Report. CITI, Montreal.

Macklovitch, E. and M.-L. Hannan (1996). Line 'Em Up: Advances in Alignment Technology and Their Impact on Translation Support Tools. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 145-156.

Marinai, E., C. Peters and E. Picchi (1991). Bilingual reference corpora: A system for parallel text retrieval. *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*. Waterloo, UW Centre for the New OED and Text Research: 63-70.

Martinez, R., J. Abaitua and A. Casillas (1998). Bitext Correspondences through Rich Mark-up. *Proceedings of COLING-ACL '98*. Montreal: 812-818.

Matsumoto, Y., H. Ishimoto, T. Utsuro and M. Nagao (1993). Structural Matching of Parallel Texts. *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics (ACL-93)*: 23-30.

Melamed, I.D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge: 184-198.

Melamed, I.D. (1996a). Automatic Construction of Clean Broad-Coverage Translation Lexicons. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 125-134.

Melamed, I.D. (1996b). A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report #96-22, Philadelphia, PA, Dept. of Computer and Information Science, University of Pennsylvania.

Melamed, I.D. (1996c). Automatic Detection of Omissions in Translations. *Proceedings from the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen: 764-769.

Melamed, I.D. (1997a). *Automatic Discovery of Non-Compositional Compounds in Parallel Data*. 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2), Providence.

Melamed, I.D. (1997b). *A Word-to-Word Model of Translational Equivalence*. 35th Conference of the Association for Computational Linguistics (ACL'97), Madrid.

Melamed, I.D. (1998a). Annotation Style Guide for the Blinker Project. Philadelphia, IRCS Technical Report #98-06, Dept. of Computer and Information Science, University of Pennsylvania.

Melamed, I.D. (1998b). Empirical Methods for MT Lexicon Construction. *Machine Translation and the Information Soup*. D. Farwell, L. Gerber and E. Hovy. Berlin, Springer-Verlag: 18-30.

Melamed, I.D. (1998c). Manual Annotation of Translational Equivalence: The Blinker Project, IRCS Technical Report #98-07, Dept. of Computer and Information Science, University of Pennsylvania.

Melamed, I.D. (1999). "Bitext Maps and Alignment via Pattern Recognition." *Computational Linguistics* 25(1): 107-130.

Merkel, M. (1992). Recurrent Patterns in Technical Documentation, Research report LiTH-IDA-R-92-31, Department of Computer and Information Science, Linköping University.

Merkel, M. (1993). When and why should translations be reused. *Papers from the XIII VAAKKI symposium*. Vaasa: 139-149.

Merkel, M. (1996). Checking Translations for Inconsistency – A Tool for the Editor. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 157-167.

Merkel, M. (1998). Consistency and variation in technical translations - a study of translators' attitudes. *Unity in Diversity. Current trends in translation studies.* Bowker, L. M. Cronin, D. Kenny and J. Pearson (eds.). Manchester, St. Jerome Publishing: 137-150.

Merkel, M. (1999). Annotation Style Guide for the PLUG Link Annotater. Linköping. PLUG report, Linköping University.

Merkel, M., B. Nilsson and L. Ahrenberg. (1994). A Phrase-Retrieval System Based on Recurrence. *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2).* Kyoto: 99-108.

Merkel, M., M. Andersson and L. Ahrenberg. (1999). The PLUG Link Annotator - Interactive Construction of Data from Parallel Corpora. *Proceedings from Parallel Corpus Symposium*, Uppsala, April 22-23, 1999, Uppsala University, under publication.

Merkel, M. and L. Ahrenberg (1999). Evaluating Word Alignment Systems. PLUG report, Linköping University.

Meyers, A., R. Yangarber, R. Grishman, C. Macleod and A. Moreno-Sandoval (1998). Deriving Transfer Rules from Dominance-Preserving Alignments. *Proceedings of Coling-ACL'98.* Montreal, Université de Montréal. II: 843-847.

Microsoft (1993). *Swedish Style Guide.* Dublin, Microsoft Worldwide Product Group Ireland.

Mitamura, T. and E.H. Nyberg 3rd (1995). Controlled English for Knowledge-Based MT: Experience with the KANT System. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation.* Leuven: 158-172.

Munday, J. (1998). "A computer-assisted approach to the analysis of translation shifts." *Meta* XLIII(4): 542-556.

Nagao, M (1984). A Framework of a mechanical translation between Japanese and English by analogy principle. A. Elithorn and R, Banerji (eds.), *Artificial and Human Intelligence*, Amsterdam, North Holland: 173-180.

Nagao, M. and S. Mori (1994). A New Method of N-gram Statistics for Large Number of *n* and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. *Proceedings from the 15th International Conference on Computational Linguistics (Coling-94).* Kyoto: 611-615.

Newmark, P. (1988). *A Textbook of Translation.* London, Prentice Hall.

Nida, E. (1964). *Toward a Science of Translating.* Leiden, E.J. Brill.

Nirenburg, S. (1987). Knowledge and choices in machine translation. *Machine translation - Theoretical and methodological issues.* S. Nirenburg, Cambridge University Press.

Norstedts (1996). *Norstedts stora engelska ordbok.* Norstedts ordböcker för PC. CD-ROM, Norstedts Förlag AB, Stockholm.

O'Brien, S. (1999). "Translation Memoryh as a linguistic resource in the Localisation Industry. A snapshot of the present and glance into the future." *The ELRA Newsletter*(April - June): 8-9.

Palmer, D.D. and M.A. Hearst (1994). Adaptive Sentence Boundary Disambiguation. *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94).* Stuttgart: 78-83.

Petti, V., I. Hesslin Rider, Y. Martinsson-Visser and A. Odeldahl (1994). *Norstedts Stora Engelsk-Svenska Ordbok*, Norstedts, Stockholm.

Platzack, C. (1983). Sex översättningar till svenska av Lewis Carrols "Alice in Wonderland". *Från språk till språk- Sjutton uppsatser om litterär översättning.* G. Engvall and R. Geijerstam. Lund, Studentlitteratur: 247-267.

Resnik, P. and I.D. Melamed (1997). Semi-Automatic Acquisition of Domain-Specific Translation Lexicons. *Proceedings of the 5th Conference on Applied Natural Language Processing.* Washington DC: 340-347.

Ridings, D. (1998). "PEDANT: Parallel Texts in Göteborg." *Lexikos* 8: 243-268.

Salkie, R. (1997). Naturalness and Contrastive Linguistics. In B. Lewandowska-Tomaszcyk and P.J. Melia (eds.). *Practical Applications in Language Corpora.* Lódz: Lódz University Press. 297-312.

Sato, S. and M. Nagao. 1990. Towards memory based translation. *Proceedings of COLING-90*, Volume 3, Helsinki: 247-252.

Schäler, R. (1994). A Practical Evaluation of an Integrated Translation Tool during a Large Scale Localisation Project. *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94).* Stuttgart: 192-193.

Shimohata, S., T. Sugio, and J. Nagata (1997). *Retrieving Collocations by Co-occurrences and Word Order Constraints.* 35th Conference of the Association for Computational Linguistics (ACL'97), Madrid: 476-481.

Sigurd, B., M. Eeg-Olofsson, C. Willners and C. Johansson. (1992). Automatic translation in specific domains (Weathra) and stock market (Stocktra, Vectra). Department of Linguistics, Lund university, Lund.

Simard, M., G.F. Foster and P. Isabelle. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation.* Montreal: 67-82.

Simard, M., G.F. Foster and F. Perrault (1993). TransSearch: A Bilingual Concordance Tool. Laval, Centre for Information Technology Innovation.

Simard, M. and P. Plamondon (1996). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*. Montreal: 135-144.

Sinclair, J., (ed.). (1987). *Looking up: An Account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London, HarperCollins.

Sinclair, J. (1995). *Collins Cobuild English Dictionary*. London, HarperCollins.

Smadja, F. (1993). "Retrieving Collocations from Text: Xtract." *Computational Linguistics* 19(1): 143-177.

Ström, A.-M. and K. Windfeldt (1991). *Klartext på NLSG*, IBM Svenska AB.

SUC (1997). *SUC 1.0 Stockholm-Umeå Corpus*. CD-ROM. Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University.

Sumita, E., K. Oi, O. Furuse, H. Iida, T. Higuchi, N. Takahashi, and H. Kitano (1993). Example-Based Machine Translation on Massively Parallel Processors. *Proceedings of the 13th International Joint Conference on Artificial Intelligence, IJCAI-93*, Chambery: 1283-1289.

Sågvall Hein, A. (1994). Preferences and Linguistic Choices in the Multra Machine Translation System. NODALIDA. *Proceedings of 9:e Nordiska Datalingvistikdagarna*. R. Eklund (ed.). Stockholm: 267-276.

Tapanainen, P. (1996). *The Constraint Grammar Parser CG-2*. Publ. 27, Dept. of General Linguistics, University of Helsinki.

Tiedemann, J. (1997). Automatic Lexicon Extraction from Aligned Bilingual Corpora. Diploma thesis. Otto-von-Guericke-Universität, Magdeburg.

Tiedemann, J. (1998). Extraction of Translation Equivalents from Parallel Corpora *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen: 120-128.

Toury, G. (1977*). Translational Norms and Literary Translation into Hebrew, 1930-1945*. Tel Aviv, The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (1980). *In Search of a Theory of Translation*. Tel Aviv, The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (1995). *Descriptive Translation Studies and beyond*. Amsterdam, John Benjamins Publishing Co.

van der Eijk, P. (1993). Automating the Acquisition of Bilingual Terminology. *Proceedings of the Sixth Conference of the European Chapter of ACL (EACL-93)*: 113-119.

van Leuven-Zwart, K.M. (1990). "Translation and Original: Similarities and Dissimilarities, II." Target 2(1): 69-95.

Vasconcellos, M. (1994). The Current State of MT Usage Or: How Do I Use Thee? Let Me Count the Ways. *The LISA Forum Newsletter*. III: 21-29.

Vermeer, H. (1978). "Ein Rahmen für eine allgemeine Translationstheorie". *Lebende Sprachen 23: 99-102*.

Véronis, J. (1998). ARCADE - Tagging guidelines for word alignment. Aix-en-Provence, Univeriteté de Provence.

Véronis, J. (1999a). From the Rosetta stone to the information society - a survey of parallel text processing. *Parallel Text Processing*. J. Véronis, Kluwer.

Véronis, J., (1999b). *Parallel Text Processing*, Kluwer Academic Press (under publication).

Véronis, J. and P. Langlais (1999). Evaluation of parallel text alignment system - The ARCADE project. To be published in *Parallel Text Processing*. J. Véronis. Berlin, Kluwer.

Wollin, L. (1981). *Svensk Latinöversättning. 1. Processen*. Lund.

Wollin, L. (1993). Translation in a European community - an old story. Stockholm. ITS Report No. 4. Institute for Interpretation and Translation Studies, University of Stockholm.

Wu, D. (1994). Aligning a Parallell English-Chinese Corpus Statistically with Lexical Criteria. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*: 80-87.

Wu, D. (1995). An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words. *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-95)*. Cambridge, Massachusetts: 244-251.

Yamamoto, M. and K.W. Church (1998). Using Suffix Arrays to Compute Term Frequency and Document Frequency for all Substrings in a Corpus. *Proceedings of the Sixth Workshop on Very Large Corpora*. E. Charniak. Montreal: 28-37.

Yang, J. and E.D. Lange (1998). SYSTRAN on AltaVista - A User Study on Real-Time Machine Translation on the Internet. *Machine Translation and the Information Soup*. D. Farwell, L. Gerber and E. Hovy. Berlin, Springer: 275-285.

Zhou, J. and P. Dapkus (1995). Automatic Suggestion of Significant Terms for a Predefined Topic. *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge: 131-147.

Department of Computer and Information Science
Linköpings universitet

Dissertations

Linköping Studies in Science and Technology

No 14 **Anders Haraldsson:** A Program Manipulation System Based on Partial Evaluation, 1977, ISBN 91-7372-144-1.

No 17 **Bengt Magnhagen:** Probability Based Verification of Time Margins in Digital Designs, 1977, ISBN 91-7372-157-3.

No 18 **Mats Cedwall:** Semantisk analys av processbeskrivningar i naturligt språk, 1977, ISBN 91-7372-168-9.

No 22 **Jaak Urmi:** A Machine Independent LISP Compiler and its Implications for Ideal Hardware, 1978, ISBN 91-7372-188-3.

No 33 **Tore Risch:** Compilation of Multiple File Queries in a Meta-Database System 1978, ISBN 91-7372-232-4.

No 51 **Erland Jungert:** Synthesizing Database Structures from a User Oriented Data Model, 1980, ISBN 91-7372-387-8.

No 54 **Sture Hägglund:** Contributions to the Development of Methods and Tools for Interactive Design of Applications Software, 1980, ISBN 91-7372-404-1.

No 55 **Pär Emanuelson:** Performance Enhancement in a Well-Structured Pattern Matcher through Partial Evaluation, 1980, ISBN 91-7372-403-3.

No 58 **Bengt Johnsson, Bertil Andersson:** The Human-Computer Interface in Commercial Systems, 1981, ISBN 91-7372-414-9.

No 69 **H. Jan Komorowski:** A Specification of an Abstract Prolog Machine and its Application to Partial Evaluation, 1981, ISBN 91-7372-479-3.

No 71 **René Reboh:** Knowledge Engineering Techniques and Tools for Expert Systems, 1981, ISBN 91-7372-489-0.

No 77 **Östen Oskarsson:** Mechanisms of Modifiability in large Software Systems, 1982, ISBN 91-7372-527-7.

No 94 **Hans Lunell:** Code Generator Writing Systems, 1983, ISBN 91-7372-652-4.

No 97 **Andrzej Lingas:** Advances in Minimum Weight Triangulation, 1983, ISBN 91-7372-660-5.

No 109 **Peter Fritzson:** Towards a Distributed Programming Environment based on Incremental Compilation,1984, ISBN 91-7372-801-2.

No 111 **Erik Tengvald:** The Design of Expert Planning Systems. An Experimental Operations Planning System for Turning, 1984, ISBN 91-7372-805-5.

No 155 **Christos Levcopoulos:** Heuristics for Minimum Decompositions of Polygons, 1987, ISBN 91-7870-133-3.

No 165 **James W. Goodwin:** A Theory and System for Non-Monotonic Reasoning, 1987, ISBN 91-7870-183-X.

No 170 **Zebo Peng:** A Formal Methodology for Automated Synthesis of VLSI Systems, 1987, ISBN 91-7870-225-9.

No 174 **Johan Fagerström:** A Paradigm and System for Design of Distributed Systems, 1988, ISBN 91-7870-301-8.

No 192 **Dimiter Driankov:** Towards a Many Valued Logic of Quantified Belief, 1988, ISBN 91-7870-374-3.

No 213 **Lin Padgham:** Non-Monotonic Inheritance for an Object Oriented Knowledge Base, 1989, ISBN 91-7870-485-5.

No 214 **Tony Larsson:** A Formal Hardware Description and Verification Method, 1989, ISBN 91-7870-517-7.

No 221 **Michael Reinfrank:** Fundamentals and Logical Foundations of Truth Maintenance, 1989, ISBN 91-7870-546-0.

No 239 **Jonas Löwgren:** Knowledge-Based Design Support and Discourse Management in User Interface Management Systems, 1991, ISBN 91-7870-720-X.

No 244 **Henrik Eriksson:** Meta-Tool Support for Knowledge Acquisition, 1991, ISBN 91-7870-746-3.

No 252 **Peter Eklund:** An Epistemic Approach to Interactive Design in Multiple Inheritance Hierarchies,1991, ISBN 91-7870-784-6.

No 258 **Patrick Doherty:** NML3 - A Non-Monotonic Formalism with Explicit Defaults, 1991, ISBN 91-7870-816-8.

No 260 **Nahid Shahmehri:** Generalized Algorithmic Debugging, 1991, ISBN 91-7870-828-1.

No 264 **Nils Dahlbäck:** Representation of Discourse-Cognitive and Computational Aspects, 1992, ISBN 91-7870-850-8.

No 265 **Ulf Nilsson:** Abstract Interpretations and Abstract Machines: Contributions to a Methodology for the Implementation of Logic Programs, 1992, ISBN 91-7870-858-3.

No 270 **Ralph Rönnquist:** Theory and Practice of Tense-bound Object References, 1992, ISBN 91-7870-873-7.

No 273 **Björn Fjellborg:** Pipeline Extraction for VLSI Data Path Synthesis, 1992, ISBN 91-7870-880-X.

No 276 **Staffan Bonnier:** A Formal Basis for Horn Clause Logic with External Polymorphic Functions, 1992, ISBN 91-7870-896-6.

No 277 **Kristian Sandahl:** Developing Knowledge Management Systems with an Active Expert Methodology, 1992, ISBN 91-7870-897-4.

No 281 **Christer Bäckström:** Computational Complexity of Reasoning about Plans, 1992, ISBN 91-7870-979-2.

No 292 **Mats Wirén:** Studies in Incremental Natural Language Analysis, 1992, ISBN 91-7871-027-8.

No 297 **Mariam Kamkar:** Interprocedural Dynamic Slicing with Applications to Debugging and Testing, 1993, ISBN 91-7871-065-0.

No 302 **Tingting Zhang:** A Study in Diagnosis Using Classification and Defaults, 1993, ISBN 91-7871-078-2.

No 312 **Arne Jönsson:** Dialogue Management for Natural Language Interfaces - An Empirical Approach, 1993, ISBN 91-7871-110-X.

No 338 **Simin Nadjm-Tehrani**: Reactive Systems in Physical Environments: Compositional Modelling and Framework for Verification, 1994, ISBN 91-7871-237-8.

No 371 **Bengt Savén:** Business Models for Decision Support and Learning. A Study of Discrete-Event Manufacturing Simulation at Asea/ABB 1968-1993, 1995, ISBN 91-7871-494-X.

No 375 **Ulf Söderman:** Conceptual Modelling of Mode Switching Physical Systems, 1995, ISBN 91-7871-516-4.

No 383 **Andreas Kågedal:** Exploiting Groundness in Logic Programs, 1995, ISBN 91-7871-538-5.

No 396 **George Fodor:** Ontological Control, Description, Identification and Recovery from Problematic Control Situations, 1995, ISBN 91-7871-603-9.

No 413 **Mikael Pettersson:** Compiling Natural Semantics, 1995, ISBN 91-7871-641-1.

No 414 **Xinli Gu:** RT Level Testability Improvement by Testability Analysis and Transformations, 1996, ISBN 91-7871-654-3.

No 416 **Hua Shu:** Distributed Default Reasoning, 1996, ISBN 91-7871-665-9.

No 429 **Jaime Villegas:** Simulation Supported Industrial Training from an Organisational Learning Perspective - Development and Evaluation of the SSIT Method, 1996, ISBN 91-7871-700-0.

No 431 **Peter Jonsson:** Studies in Action Planning: Algorithms and Complexity, 1996, ISBN 91-7871-704-3.

No 437 **Johan Boye:** Directional Types in Logic Programming, 1996, ISBN 91-7871-725-6.

No 439 **Cecilia Sjöberg:** Activities, Voices and Arenas: Participatory Design in Practice, 1996, ISBN 91-7871-728-0.

No 448 **Patrick Lambrix:** Part-Whole Reasoning in Description Logics, 1996, ISBN 91-7871-820-1.

No 452 **Kjell Orsborn:** On Extensible and Object-Relational Database Technology for Finite Element Analysis Applications, 1996, ISBN 91-7871-827-9.

No 459 **Olof Johansson:** Development Environments for Complex Product Models, 1996, ISBN 91-7871-855-4.

No 461 **Lena Strömbäck:** User-Defined Constructions in Unification-Based Formalisms,1997, ISBN 91-7871-857-0.

No 462 **Lars Degerstedt:** Tabulation-based Logic Programming: A Multi-Level View of Query Answering, 1996, ISBN 91-7871-858-9.

No 475 **Fredrik Nilsson:** Strategi och ekonomisk styrning - En studie av hur ekonomiska styrsystem utformas och används efter företagsförvärv, 1997, ISBN 91-7871-914-3.

No 480 **Mikael Lindvall:** An Empirical Study of Requirements-Driven Impact Analysis in Object-Oriented Software Evolution, 1997, ISBN 91-7871-927-5.

No 485 **Göran Forslund**: Opinion-Based Systems: The Cooperative Perspective on Knowledge-Based Decision Support, 1997, ISBN 91-7871-938-0.

No 494 **Martin Sköld**: Active Database Management Systems for Monitoring and Control, 1997, ISBN 91-7219-002-7.

No 495 **Hans Olsén**: Automatic Verification of Petri Nets in a CLP framework, 1997, ISBN 91-7219-011-6.

No 498 **Thomas Drakengren:** Algorithms and Complexity for Temporal and Spatial Formalisms, 1997, ISBN 91-7219-019-1.

No 502 **Jakob Axelsson:** Analysis and Synthesis of Heterogeneous Real-Time Systems, 1997, ISBN 91-7219-035-3.

No 503 **Johan Ringström:** Compiler Generation for Data-Parallel Programming Langugaes from Two-Level Semantics Specifications, 1997, ISBN 91-7219-045-0.

No 512 **Anna Moberg:** Närhet och distans - Studier av kommunikationsmmönster i satellitkontor och flexibla kontor, 1997, ISBN 91-7219-119-8.

No 520 **Mikael Ronström:** Design and Modelling of a Parallel Data Server for Telecom Applications, 1998, ISBN 91-7219-169-4.

No 522 **Niclas Ohlsson:** Towards Effective Fault Prevention - An Empirical Study in Software Engineering, 1998, ISBN 91-7219-176-7.

No 526 **Joachim Karlsson:** A Systematic Approach for Prioritizing Software Requirements, 1998, ISBN 91-7219-184-8.

No 530 **Henrik Nilsson:** Declarative Debugging for Lazy Functional Languages, 1998, ISBN 91-7219-197-x.

No 555 **Jonas Hallberg:** Timing Issues in High-Level Synthesis,1998, ISBN 91-7219-369-7.

No 561 **Ling Lin:** Management of 1-D Sequence Data - From Discrete to Continuous, 1999, ISBN 91-7219-402-2.

No 563 **Eva L Ragnemalm:** Student Modelling based on Collaborative Dialogue with a Learning Companion, 1999, ISBN 91-7219-412-X.

No 567 **Jörgen Lindström:** Does Distance matter? On geographical dispersion in organisations, 1999, ISBN 91-7219-439-1.

No 582 **Vanja Josifovski:** Design, Implementation and Evaluation of a Distributed Mediator System for Data Integration, 1999, ISBN 91-7219-482-0.

No 589 **Rita Kovordányi**: Modeling and Simulating Inhibitory Mechanisms in Mental Image Reinterpretation - Towards Cooperative Human-Computer Creativity, 1999, ISBN 91-7219-506-1.

No 592 **Mikael Ericsson:** Supporting the Use of Design Knowledge - An Assessment of Commenting Agents, 1999, ISBN 91-7219-532-0.

No 593 **Lars Karlsson:** Actions, Interactions and Narratives, 1999, ISBN 91-7219-534-7.

No 594 **C. G. Mikael Johansson:** Social and Organizational Aspects of Requirements Engineering Methods - A practice-oriented approach, 1999, ISBN 91-7219-541-X.

No 595 **Jörgen Hansson:** Value-Driven Multi-Class Overload Management in Real-Time Database Systems, 1999, ISBN 91-7219-542-8.

No 596 **Niklas Hallberg:** Incorporating User Values in the Design of Information Systems and Services in the Public Sector: A Methods Approach, 1999, ISBN 91-7219-543-6.

No 597 **Vivian Vimarlund:** An Economic Perspective on the Analysis of Impacts of Information Technology: From Case Studies in Health-Care towards General Models and Theories, 1999, ISBN 91-7219-544-4.

No 598 **Johan Jenvald:** Methods and Tools in Computer-Supported Taskforce Training, 1999, ISBN 91-7219-547-9.

No 607 **Magnus Merkel:** Understanding and enhancing translation by parallel text processing, 1999, ISBN 91-7219-614-9.

Linköping Studies in Information Science

No 1 **Karin Axelsson:** Metodisk systemstrukturering - att skapa samstämmighet mellan informationssystemarkitektur och verksamhet, 1998. ISBN-9172-19-296-8.

No 2 **Stefan Cronholm:** Metodverktyg och användbarhet - en studie av datorstödd metodbaserad systemutveckling, 1998. ISBN-9172-19-299-2.