Linköping Electronic Articles in
Computer and Information Science
Vol. 4(1999): nr 2

Automatic Processing of Parallel Corpora: A Swedish Perspective

Lars Ahrenberg

Magnus Merkel

Department of Computer and Information Science
Linköping University
Linköping, Sweden

Daniel Ridings

Department of Swedish Language
Göteborg University
Göteborg, Sweden

Anna Sågvall Hein

Jörg Tiedemann

Department of Linguistics
Uppsala University
Uppsala, Sweden

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/1999/002/

Published on December 22, 1999 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

©*1999 Lars Ahrenberg Magnus Merkel Daniel Ridings Anna Sågvall Hein Jörg Tiedemann*
Typeset by the authors using LaTeX
Formatted using étendu *style*

Recommended citation:
<Authors>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 4(1999): nr 2. http://www.ep.liu.se/ea/cis/1999/002/. *December 22, 1999.*

This URL will also contain a link to the authors' home pages.

The publishers will keep this article on-line on the Internet (or its possible replacement network in the future) for a period of 25 years from the date of publication, barring exceptional circumstances as described separately.

The on-line availability of the article implies a permanent permission for anyone to read the article on-line, to print out single copies of it, and to use it unchanged for any non-commercial research and educational purpose, including making copies for classroom use. This permission can not be revoked by subsequent transfers of copyright. All other uses of the article are conditional on the consent of the copyright owners.

The publication of the article on the date stated above included also the production of a limited number of copies on paper, which were archived in Swedish university libraries like all other written works published in Sweden. The publisher has taken technical and administrative measures to assure that the on-line version of the article will be permanently accessible using the URL stated above, unchanged, and permanently equal to the archived printed copies at least until the expiration of the publication period.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its WWW home page: http://www.ep.liu.se/
or by conventional mail to the address stated above.

Abstract

As empirical methods have come to the fore in multilingual language technology and translation studies, the processing of parallel texts and parallel corpora have become a major research area in computational linguistics. In this article we review the state of the art in alignment and data extraction techniques for parallel texts, and give an overview of current work in Sweden in this area. In a final section, we summarize the results achieved so far and make some proposals for future research.

This report has been produced as part of the project Datagenerering från parallellkorpusar *financed by The Swedish Council for Research in the Humanities and Social Sciences (HSFR) and The Swedish National Board for Industrial and Technical Development (NUTEK) within their joint research programme in language technology, grant F0293/97.*

1 Introduction

In recent years much progress has been made in computational linguistics in developing tools and methods for the alignment of source texts with their translations, and for extracting translation data from the resulting parallel text.

The purpose of this paper is to review the state of the art in automatic processing of parallel texts and describe current Swedish work in the area. In a final section, we discuss the results achieved so far and make some proposals for future research.

1.1 Why parallel corpora

While machine translation has been conceived of since at least 1947 [Hutchins:86] and monolingual corpora have been created and processed by computers since the sixties [Leech93], the creation and processing of bilingual and multilingual corpora on a larger scale took a much longer time to get started. Why this was so is an issue that we will not delve into, but apart from the computational problems involved, part of the explanation can probably be found in the dominating theoretical orientation of both translation researchers and computational linguists interested in machine translation.

Towards the end of the eighties there was a marked change, however. In machine translation new models were proposed, such as example-based translation [SN90] and, in particular, the statistical models of [BCDD90]. The nineties have seen systems for computer-aided translation based on translation databases (translation memories) make a heavy impact on the market, and such databases have been used also for other purposes, e.g. for customizing traditional MT-systems, in bilingual lexicography, terminology, contrastive linguistics and in translator's training. In fact, a large part of the early work focus on tools and data for the lexicographer [KT96, CWR90].

Parallel texts provide the basis for many applications and tools. One of the most important is the multilingual concordance, where words and phrases are shown in the context of both the sentences in which they occur, and the translations of these sentences. This is of great use to lexicographers as well as terminologists, translators and language learners. They are also used for the generation of potential dictionary entries and equivalent terms. Another possibility is for use with dynamic dictionaries, i.e. dictionaries where the information about entries is adapted to the specific properties of a given text type or domain, something which is of special importance in language technology projects. Other uses are in translator's workbenches where they provide an initial translation memory, and for the testing and development of machine translation systems of various kinds.

1.2 Basic concepts

By a parallel text we understand a source text and its translation into one or more target languages. In the special case where there is just one target language we speak of the parallel text as a *bitext*. An *alignment* is a mapping or linking of the two halves of a bitext, that associates segments of one half with corresponding segments of the other half.

In the last decade several projects aiming at the establishment, preparation and maintenance of parallel corpora as a basis for linguistic research and development have been initiated. In particular, much effort has been devoted to questions concerning the retrieval of multilingual information and translation data from parallel corpora.

The basic steps in the creation of parallel texts are the following:

- **Capture**, the collection of computer-generated texts in the same environment, possibly including conversion from paper format to a computer text.
- **Cleaning up**, the correction of errors and removal of formatting information.
- **Segmentation**, proper identification of the segments of interest, e.g. paragraphs, sentences and words.
- **Annotation**, mark-up of the texts as wholes and of the segments of interest.
- **Alignment**, linking of segments of one text with corresponding segments of the other texts.

Given an aligned parallel text it can then be processed in various ways to obtain interesting translation data, e.g. lexical and terminological correspondences, contrastive phraseology and valency data, and data on variation in the translations of given words and phrases.

In this report we will focus on the alignment and subsequent retrieval of translation equivalents at different levels, in particular sentences and words, and on the development of search tools for aligned texts (bilingual and multilingual concordances).

Section 2 gives a brief review of current methods for creating and processing parallel text. Section 3 reports on recent and ongoing research in Sweden in the area. Finally, Section 4, gives an overview of some important research issues.

2 Current methods

This section gives a brief presentation of current methods for processing parallel text. The presentation is based on [Tie97].

2.1 Sentence alignment of parallel text

Sentence alignment is a fundamental step in preparing parallel corpora for further investigation of translation relations and the retrieval of translation equivalents. The task of sentence alignment is not trivial. Problems are due to various operations that the translator may perform during translation resulting in removed, inserted, split, or combined sentences. Consequently, a sentence alignment program has to account for 1-0 (sentence deletion), 0-1 (sentence insertion), x-1 (sentence combination), 1-x (sentence splitting), and x-y (multiple source language sentences differently split into multiple sentences in the target language). The most frequent alignment type though will be 1-1. The following table contains two examples from the Scania corpus (see 3.3.2) with alignments from Swedish to English.

Swedish	English
Vanligaste enheten är kg/dm^3 (vattens densitet är 1 kg/dm^3).	The most common unit is kg/dm^3. The density of water is 1kg/dm^3.
Specifik vikt är en annan benämning för densitet	Specific gravity is another term for density

Several methods for automatic sentence alignment have been developed in the past. The next sections will introduce the most important current ones.

Alignment based on sentence length

In [GC93] William Gale and Ken Church propose a simple statistical algorithm for aligning parallel texts. The article includes a core implementation of the proposed algorithm in C. The approach is based on the assumption that corresponding text parts tend to have a similar length in terms of characters.

The application of GC-align presupposes that the text to be aligned has been divided into *hard regions* (e.g. paragraphs, sections). The hard regions are then analysed in terms of sentence lengths. Sentence length is measured in terms of characters although word length (number of words per sentence) could be used. In practical tests it turned out that using the character length yielded the highest precision.

GC-align is able to recognize 1-1, 1-0, 0-1, 2-1, 1-2, and 2-2 sentence alignments. The algorithm computes a probabilistic score for every alignment type within the hard region. Based on these probabilities the program computes the optimal set of alignments using a dynamic programming algorithm [GC93].

Although the GC-alignment approach uses very simple assumptions it yields highly correct sentence alignments in practical applications. In [GC93] the authors claim an average precision of about 96% based on performing the algorithm to a trilingual corpus in English,

French, and German of economic reports issued by the Union Bank of Switzerland.

Cognates and anchor words

Char_align [Chu93] is a statistical program for aligning parallel text that was developed at the AT&T Bell laboratories. It aligns texts at the character level applying the cognate approach proposed by Simard, Foster, and Isabelle [SFI92]. This approach is based on the existence of cognates between cross-language token pairs. In [SFI92] the authors propose to use these cognates to improve a length-based sentence-alignment method by defining a 'level of cognateness' as follows [SFI92]:

$$\gamma = \frac{c}{(n+m)/2} \tag{1}$$

Here, c is the maximum number of cognates (after matching the target language token to the source language token) in the current text unit (sentence), n is the number of tokens in the source language and m is the number of tokens in the target language.

Char_align uses identical 4-grams to find an alignment path between source and target language text. For this purpose the program uses a 'dotplot calculation' [CH93]. If there is a 4-gram at position x in the source file, and an identical 4-gram at position y in the target file, the corresponding flag will be set in a two-dimensional xy-array.

In the final step, the best alignment path between these 'dots' has to be identified. For this purpose, the following heuristic is used. The path with the largest average weight, i.e. the ratio between the sum of the weights along the path and its length, will be considered as the best alignment path [Chu93]. This path is the basic criterion for aligning the bilingual text.

The Scandinavian Project of Contrastive Corpus Studies 3.4, uses an alignment program by Knut Hofland [Hofl95]. This program uses a more linguistic and language specific approach. By using cognates as well as lists of anchor words they find the words in the text that are most likely to be translated in a predictable way and therefore can point to anchor points for the alignment of two texts.

The K-vec Method

The K-vec algorithm is another statistical approach to bilingual text alignment and it was developed by Pascale Fung [FC94] at the Computer Science Department at Columbia University, New York.

The first step in the application of this method is to extract lexicon candidates by looking for similarities in the distribution of source and target language words. For this purpose the bilingual text is split into K pieces. After that, K-dimensional binary vectors are created for the words in the source and the target language. If a specific text piece contains the source language word (resp. the target language

word), a corresponding flag in the vector is set. Then, statistical methods can be used to find the similarities between these words.

The K-vec method uses a mutual information score which is defined as follows:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2}$$

$P(x,y)$ is the probability that the words x and y occur in corresponding pieces, and $P(x)$ and $P(y)$ are the probabilities that x respective y occur in the text.

The probabilities can be estimated by using absolute frequency numbers. $P(x,y)$ will be $\frac{freq(x,y)}{K}$, $P(x)$ and $P(y)$ will be $\frac{freq(x)}{K}$ and $\frac{freq(y)}{K}$, where $freq(x,y)$ is the frequency that x and y occur together and $freq(x)$ and $freq(y)$ are the global numbers of the occurrence of x and y. Using this mutual information measure, translation candidates can be ranked, and the most likely pair will be chosen as the corresponding pair. These pairs are then used as reference points to align the parallel text corpora. The alignment task may be perfomed by a separate alignment algorithm, e.g. the one proposed in [Chu93].

Problems with mutual information score arise with low frequency words [FC94]. The t-score [CGHH91, FC94] is used to filter out insignificant values.

The K-vec algorithm was applied to the Canadian Hansards corpus[1]. The resulting estimation of a bilingual dictionary could be used in sentence alignment.

DK-vec

Based on the K-vec method, Pascale Fung and Kathleen McKeown from Columbia University, New York, developed a new algorithm for aligning noisy parallel corpora. By a noisy parallel corpus we understand parallel text documents with missing or incomplete segments in some languages. Basically, they can be considered as parallel texts but they cannot be processed on a segment by segment basis. The DK-vec approach uses a matching algorithm called Dynamic Time Warping [FM94].

In the DK-vec algorithm, the distances between the occurrences of the source and the target language words are stored in so-called recency vectors [FM94]. Due to different word frequencies the recency vectors differ in size. They are used by the Dynamic Time Warping algorithm to find matches between the source and the target language.

The DK-vec algorithm starts by computing the recency vectors for each source and target language word. Next, translation candidates have to be located. For this purpose, all candidate pairs which occur for the first time in the second half of the text are filtered out. Furthermore, all pairs with one vector less than half of the length of

[1] Canadian parliamentary proceedings.

the others are removed. For the remaining pairs the absolute difference between their vectors is computed by means of Dynamic Time Warping (see further [FM94]). In a final step, the word pairs are sorted by absolute difference and closely correlated word pairs are identified. As in the K-vec method these pairs are used as reference points to align the parallel text.

Bitext Mapping - SIMR

A method for the mapping of bilingual text correspondences called the Smooth Injective Map Recognizer (SIMR) was developed at the University of Pennsylvania in Philadelphia by Dan Melamed [Mel96]. Like char_align it uses cognates in the source and target language text to align bilingual corpora at character level.

A bilingual text is seen as a two-dimensional 'bitext space' [Mel96] with the character position of the source language text and the character position of the target language text as axes. Starting with the origin of the bitext space, the algorithm searches for so-called 'true points of correspondence' (TPC's) [Mel96] by expanding the search space in the direction of x and y in ascending order. These new TPC's can be identified by the use of machine-readable dictionaries, cognate-based matching algorithms, e.g. the longest common subsequence ratio (LCSR) [Mel96], or other matching algorithms.

If a new TPC is discovered, and its position does not have the same x or the same y value in the bitext space as the former origin, the new origin is based on the newly discovered TPC for the next search step [Mel96].

After discovering the TPC's, an alignment path has to be chosen. For this purpose the most linear and constant path between the origin and the terminus of the bitext space is chosen by filtering path candidates in several steps (see [Mel96]).

The obtained alignment path can be used to align sentences of the bilingual text.

2.2 Storing and accessing parallel corpus data

Storing parallel corpus data

The required format for storing parallel corpus data is highly dependent on processing level. Here it is useful to distinguish between plain parallel text (ASCII), aligned parallel text (e.g. sentence aligned), tokenized text, and tagged text. Combinations of the last three types may also occur.

Standard Generalized Markup Language (SGML) is the most commonly used format for storing corpus data. It is well suited for representing hierarchical information about text. Usually, monolingual documents are encoded in SGML, and pointer structures may be used to establish connections between parallel documents. Hereby, each monolingual part will be stored coherently and may be processed

```
94X:1:2:1:2:
    (1) Specialverktyg
    (2) Special tools
 2X:4:8:4:8:
    (1) Fyll hyttippumpen med hydraulolja.
    (2) Fill the cab tilt pump with hydraulic fluid.
```

Figure 1: An example of sentence aligned bilingual corpus data from the Scania corpus.

separately. For the extraction of multilingual information, encoded links have to be analysed.

A standardized encoding scheme for SGML was defined by the Text Encoding Initiative (TEI). In the TEI guidelines [SMB94] a standard set of tags is proposed to make it easier to exchange documents between different platforms and systems. The TEI guidelines contain a complex scheme definition. In most cases it is sufficient to use only a subset of the complete tag set. An example of such a standardized subset is TEI Lite which was defined by the TEI.

A main problem with the use of SGML like structures is the enormously increasing size of the encoded texts.

There are, however, alternative formats for storing parallel text. For instance, aligned sentences from different languages may be stored together in one single file. An example of such a format is presented in figure 1.

Another possibility for storing text corpora is to store actual data (ASCII text) and structural information about the text in different locations. The implementation of such a strategy is based on the establishment of data files that contain pointers to the actual text data by referring to their positions in the text file. An approach using this principle is the TIPSTER architecture [TIP97], which uses spans (start and end byte position in a plain ASCII text) to store attribute values for specific parts of the text. An extension of this architecture for parallel texts is possible and will be specified in phase III of the Tipster project [TIP97].

A general problem with this format concerns the updating of the text. A simple insertion of one word in the text implies a new calculation of all pointers to following entities, or all annotations of entities after the inserted word will be wrong. Another problem is caused by the fact that different attributes are saved in different annotation collections. A query for several attributes may therefore be complicated.

Accessing parallel corpora

Linguistic information retrieval is the main reason for compiling and storing parallel corpora, and there is a demand for efficient tools for processing and searching such corpora.

Tools for processing parallel text corpora are highly dependent on the internal format in which the information about the text is stored. As regards SGML encoded text, tools for querying and extracting partial information from a hierarchically structured text remain to be developed. A general problem with SGML encoded corpora is their size.

A solution to the size problem may be to make a split between the text corpus as such and the linguistic information about the individual lexical units. Such a strategy may be implemented by means of a lexical database, the entries of which are linked to the word occurrences of the text. Corpus accessing tools for processing such data should provide the same kind of functionality as if the data would have been stored in one single file (transparent views). A problem with such a "split" strategy may arise with the use of different architectures for the different data modules (lexical database, textual corpus data). An advantage, on the other hand, will be faster access to the lexical data due to efficient storage in an external database. Further, the size of the text corpus will be reduced because linguistic data don't have to be stored redundantly for each lexical unit; an additional positive side-effect will be faster access.

2.3 Extraction of translation equivalents

Sentence alignment and dictionary generation go hand in hand. Some of the methods for sentence alignment described above, like the K-vec method, actually start by extracting candidate translation equivalents. Others make crucial use of an initial dictionary of some sort. Conversely, given a bitext that is correctly sentence aligned, you are in a better position to extract lexical translation data.

In this section we review some methods for extraction of lexical translation data, starting by word correspondences and then considering terminology and collocations.

Termight

The Termight system is directed towards bilingual lexicon creation. It was developed at the AT&T Bell laboratories by Ido Dagan and Ken Church, and it provides a semiautomatic tool for the identification of technical terms and their translations [DC94]. The system is based on part-of-speech tagging and word alignment. Word alignment is handled by means of *word_align* [DCG93], a bilingual alignment program for noisy texts [DC94]. Word_align is based on *char_align* [Chu93] which was introduced in section 2.1. Char_align works at the character level, and it uses cognates between the two languages (identical

4-grams). Therefore, it is restricted to historically related languages with the same set of characters. From these word alignments, translation candidates can be extracted. Termight sorts the candidates by frequency and provides the user with a concordance tool for manual filtering of the data.

The filtering approach for rating candidate word pairs

An approach developed by I. Dan Melamed at the University of Pennsylvania in Philadelphia is directed towards the automatic evaluation of lexicon by applying several filters [Mel95]. The filters use external knowledge sources and heuristics.

First, all the source language words and all the target language words of a sentence alignment pair are combined into word pairs. Then filters are applied in cascades to find the N-best (e.g. 7-best) translations among the translation candidates.

Four different filters are used: part of speech filters, machine-readable bilingual dictionary filters, cognate filters, and word alignment filters.

The part of speech filter removes every translation candidate which differs with regard to part of speech from the source language word.

The second filter uses machine-readable bilingual dictionaries (MRBD). If a translation candidate appears in the MRBD, all pairs with the same source language word and a different target language word, and all pairs with the same target language word and a different source language word occurring in the same sentence pair are removed [Mel95]. In other words, the translation of the source language word from the MRBD is assumed to be correct and all the other target language words from the same sentence alignment pair are disregarded as translation equivalents.

Cognate filters are based on the assumption that there are similarities between the source language word and its translation in related languages. In order to rank the translation candidates by level of 'cognateness' Melamed uses the Longest Common Subsequence Ratio [Mel95]. The LCSR algorithm counts the number of letters in the longest (not necessarily contiguous) common subsequence and divides this value by the length of the longer string [Mel95].

Word alignment filters assume that in related languages information is expressed in a similar word order. The heuristic here is that crossing alignments are not very probable. In other words, if a source language word s is aligned to a target language word t, all source words before s correspond more likely to the target words before t and the source words after s correspond more likely to the words after t.

This filter can be combined with the cognate filter and the part-of-speech filter.

These automatic evaluation methods can be used to prepare bilingual dictionaries for human evaluation. (It should be noted that only 1:1 word pairs are compiled).

Some experiments were made with the Canadian Hansards corpus. The precision for the single best translations was about 52% [Mel95].

Translating Collocations - Champollion

The Champollion System [SMH96] developed by Frank Smadja, Vasileios Hatzivassiloglou, and Kathleen R. McKeown at Columbia University, New York, focuses on the identification of collocations in text corpora and the automatic identification of corresponding translations for a given parallel bilingual corpus. The goal is to compile lexical data above the word level from parallel text.

The first task is to identify monolingual collocations. Here collocations are simply defined as word sequences which occur with a higher frequency than other word sequences.

A statistical collocation compiler called Xtract is used to identify collocations. It distinguishes between 'fixed' and 'flexible' collocations. Fixed collocations are frozen phrases without intervening elements, such as 'United States'. Flexible collocations can be interrupted by other words, or the word order may change. An example of a flexible collocation is 'make decision' which may appear as 'make a decision' or 'decisions to make' [Sma93].

A precondition for the further processing is a sentence alignment of the parallel corpus.

In the following processing steps, Champollion uses the Dice coefficient [Dic45] based on absolute word frequency, which is defined as follows:

$$Dice(x, y) \approx \frac{2f_{xy}}{f_x + f_y} \tag{3}$$

f_{xy} represents in this case the absolute frequencies of x and y occurring together, and f_x and f_y are absolute single word frequencies.

Using identified source language collocations, Champollion processes the aligned data in an iterative process to find the best target language phrase (according to the statistical measure). For this purpose, empirically chosen thresholds are used to filter translation candidates. First, Champollion considers target language phrases as sets of words and in the last stage, word order and collocation type are identified by an analysis of the corresponding target language sentences.

Champollion was tested with the English/French Hansards corpus (Canadian parliamentary proceedings). Three different sets of collocations were retrieved from the English version of the corpus. The evaluation of the results yielded an average value of 73% precision [SMH96].

Acquisition of Bilingual Terminology

Another statistical approach to the extraction of translation equivalents has been developed at the Digital Equipment Corporation by Pim van der Eijk. This method concentrates on the identification of noun phrase correlates from a previously aligned and tagged parallel corpus.

In a preprocessing step the corpus is sentence aligned and tagged with part-of-speech tags.

For the identification of noun phrases, a simple pattern matching algorithm is used. According to this algorithm a noun phrase is simply a sequence of zero or more adjectives followed by one or more nouns [vdE93].

$$np \rightarrow w_a^* w_n^+$$

The statistical method for finding correlates is based on the following assumption: the translation equivalent is more frequent in the subset of the target language sentences, which are aligned to the source language sentences (containing the source language term under consideration), than in the entire target language text [vdE93].

The system calculates a 'local' frequency (the frequency of the target language term candidate in the subset of the target language sentences aligned to the source language sentences containing the term under consideration) and a 'global' frequency for the target language terms. The following ratio is used to measure the correlation [vdE93].

$$\frac{f_{local}(target|source)}{f_{global}(target)} \tag{4}$$

Problems with using this score appear with low frequency words. Therefore, all target language terms with a local frequency below a certain threshold are removed [vdE93].

To improve the results, van der Eijk proposes a position-sensitive score reduction. The scores will decrease proportionally to the distance to the estimated position in the target segment [vdE93]. This method yielded remarkably improved results [vdE93].

Several experiments were made with word-based, noun-based, and phrase-based methods, respectively, yielding a precision between 33% and 77%. The best results were achieved with the word- and noun-based methods with position sensitivity, i.e. a precision of 77% with a recall of 50% [vdE93].

BICORD

Judith Klavans and Evelyn Tzoukermann concentrate on combining lexical data from bilingual machine-readable dictionaries with bilingual corpus information. In [KT96] they describe a system, *BICORD*, that supports the extension of lexicon entries in a bilingual machine-readable dictionary with information extracted from the bilingual Canadian Hansards corpus. In explaining their theoretical basis the

authors focus on the treatment of verbs of movement. They combine linguistic and statistical methods to enhance the lexicon entries and to build a useful lexical database. In support of the process, statistical part-of-speech tagging and structural analyses are used.

Other methods

A few other approaches to automatic lexicon extraction which will be mentioned here very briefly.

At the IBM research laboratories, a stochastic language system, *Candide*, was developed which uses speech recognition techniques [BDDM93]. The system considers word groups as well as single words.

Another noun phrase recognizer was developed by Kupiec [Kup93], which is based on a sentence-aligned and part-of-speech-tagged bilingual corpus. The noun phrases are recognized separately for each language and mapped to each other in an iterative process.

The last approach to be mentioned here is implemented in a system developed by Dekai Wu and Xuanyuin Xia [WX94]. For the extraction of single word translations from a sentence aligned parallel corpus, the authors use an 'estimation-maximization' technique with additional significance filtering [SMH96].

2.3.1 Available systems and corpora

This section gives some information on generally available resources for obtaining and working with parallel corpora. For more detail about Swedish resources the reader is referred to chapter 3.

Programs for alignment

As already mentioned the Gale&Church article [GC93] about sentence alignment also includes the code for such a program. This article combined with the code has been the base for many alignment programs, including those developed in Göteborg, Linköping and Uppsala. Others who have been using the method for their own alignment programs is the Lingua Parallel Concordancing project, involving partners from Denmark, Germany, Greece, Italy, England and France [BR95].

The commercial systems dealing with parallel texts also have alignment programs, such as for example the Winalign system from TRADOS. This and other similar commercial alignment programs can be purchased in combination with translator's workbenches and are primarily designed for the creation of an initial translation memory.

Corpus access tools

Here, three corpus tools will be introduced briefly.

LT NSL is a set of SGML/XML processing tools including a developer's tool-kit, and a C-based API [LTN97], which was developed at the University of Edinburgh. The set contains several general tools for processing normalized SGML files. Most interesting are the query tools 'sggrep' and 'sgrpg'. They support complex queries to every kind of normalized SGML and print resulting data to STDOUT. Furthermore, the API provides the possibility of implementing new tools using the SGML access functions of LT NSL. The advantage of using this toolkit is the independence of fixed structures and restricted functionalities. The disadvantage is the programming effort to develop specialized data management tools. Furthermore, the access speed might be a problem.

Another toolbox for querying text corpora, the IMS corpus toolbox, was developed at Stuttgart University [IMS97]. It is a set of tools for administering, indexing, and querying large text corpora. It contains two modules, a command line interface for running complex queries using a query language called 'cqp', and a graphical interface for XWindow systems called 'XKwic' with additional features like concordance generation and sorting functions. Right now there is no SGML support in the toolbox. Data are stored as sets of positional annotations such as words with associated lemma, part of speech etc., and structural annotations such as sentence boundaries. For further information, see the project home page in the WWW [IMS97].

Another collection of corpus tools was developed by the European Union sponsored MULTEXT project [MUL97]. The collection includes a series of corpus accessing and manipulating tools using the encoding and software specifications made in the project. They include SGML encoded corpus data. Among the tools there is an SGML query language interpreter and an SGML software API.

There are other possibilities of processing large parallel text corpora. The development of specific tools is highly dependent on the format that is used for storing the data and, of course, on the task at hand. One possibility, for instance, may be the use of standard query tools and programming languages with efficient text processing functionality such as Perl. The disadvantage with using a standard system is the high effort that is required to adjust the software to the specific needs of the task or to implement additional software. The disadvantage with using specialized linguistic software is the restriction of its functionality. Mostly, it is not possible to extend fixed software solutions with additional functionality.

Available parallel corpora

As the interest in parallel corpora have increased dramatically, several research projects have made efforts to collect sizeable parallel corpora in recent years. In spite of these efforts there is no abundance of available parallel corpora, even for research purposes, as distribution of the texts is conditioned by copyright laws.

A parallel corpus available for free is the one collected within the CRATER project (MLAP93/20) [MWSS97]. Others can be obtained for limited sums by purchasing the CD-ROMs created within the Eureopean Corpus Initiative. Other EU-funded projects working on parallel corpora, such as Lingua and Parole, are expected to make their corpora available via ELRA, the European Language Resources Association [2]. The Hansard Corpus of parallel English and French, covering debates in the Canadian parliament, can be obtained through the Linguistic Data Consortium [3], which also sells a parallel English-French-Spanish corpus of documents from the United Nations.

3 Current Swedish research on processing parallel corpora

On-going Swedish research on automatic processing of parallel corpora is, basically, located to the universities. Below we present on-going activities site by site with regard to projects, corpora, methods and software. In an additional section we present some projects that are also creating and/or using parallel corpora, but with less emphasis on processing.

3.1 Göteborg University

A small project dealing with parallel texts was initiated in the autumn of 1995 by Pernilla Danielsson and Daniel Ridings under the working name of PEDANT. The original plans are set out in [DR96c]. The work was a private research initiative and only received support from the Faculty of Arts within the framework of Språkbanken in 1997. The original report has been adjusted and expanded in [Ridings98], where many of the technical details for the points only touched upon in this report can be found.

Much of the earliest work was intentially geared towards awakening an interest for the work in other departments of modern languages at the faculty and the translator training program that was being planned. For this reason the main language pairs became Swedish in combination with English, French and German. All alignment work is done in such a way that Swedish is always one of the languages. No work has been done between English and French, for example. There is a substantial amount of material in other languages, Italian and Spanish, but with a couple of reservations, most remains in a raw, unprocessed state.

The first two years, 1996–1997, were dedicated to building up a substantial collection of texts from the socio-political domain, mostly

[2] http://www.icp.grenet.fr/ELRA/home.html
[3] http://www.ldc.upenn.edu/ldc/

texts from the European Union and a small percentage from the industrial sector with material from local industries. This was intended for the translator training program which had the express policy to concentrate on the needs that were created when Sweden joined the European Union. This effort had the side-effect of awakening an interest in parallel texts at the departments involved and with time approximately 600,000 words (source language) were collected from novels in Swedish-Italian and Swedish-French. The literary material is the working material for one advanced level student and one graduate student in French and Italian respectively at the Department of Romance Languages.

The text collection is growing constantly and at present consists of approximately 750,000 words (per language) in the socio-political domain covering Swedish, English, French in German in aligned form and the slightly less in Italian and Spanish in raw format.

3.1.1 Alignment

The texts are first prepared by processing them with locally developed tokenizers for each language involved. Sentence segmentation is performed during the tokenization process. The tokenizers have been steadily improved over time and perform quite well on technical and legal texts. This is a crucial step since it ensures that the corpus and database format can be mapped to each other on a one-to-one basis.

The texts are aligned using a version of Gale and Church's program (see 2.1 and [GC93]). It has proved to be a satisfactory method. The success rate has improved with time, but this is due more to good quality tokenized text than to tweaking the alignment program. It covers the following alignment pairs: 1-1, 2-1, 2-2 and 1-0. It fails most often in the case of 3-1 alignments since it cannot distinguish a 2-1 relationship followed by a 1-0 relationship from 3-1 relationships. 1-0 relationships occur so often in texts from the European Union that it was decided not to introduce a 3-1 possibility at the cost of 1-0.

3.1.2 Storage

The output of the aligner has been customized to fit in with the SGML and relational database environment being used in Gothenburg. This is described in more detail in [Ridings98] and will only be outlined here.

It was decided to use SGML to its full extent for the corpus format. This was motivated by the fact that Gothenburg was responsible for the SGML corpus format for the whole LE2-4017 PAROLE project. This experience kept Gothenburg at the very forefront of developments concerning SGML and corpus annotation.

Each language is stored as a TEI corpus document. In principle that means that there is a separate corpus for each language includ-

ed in the collection. A well-defined system of assigning attributes to the SGML elements was designed and reused as various keys in the tables of the relational database. The annotation is kept to a minimum with regards to document grammar. The only elements used extensively in the individual corpora, besides those in the teiheaders, are the elements for `TEXT`, `BODY`, `P`, `S` and `W`. Others are allowed, but have not been felt to be necessary for the requirement for the work being done with the texts. Few elements are used, but the attribute values assigned to each element are rich and unique. This allows subsequent processing to always know exactly where the data in question originates from, down to the level of the token, the base level annotation.

Besides a collection of monolingual corpora, an alignment corpus is created for each language pair, ie. Swedish-English, Swedish-German, Swedish-French and so on. These alignment corpora contain no textual data, but only SGML elements with information that refers back into the relevant monolingual corpora. When these alignment corpora are processed by the SGML tools being used by the project, the information that the pointers point to is retrieved from the individual monolingual corpora and alignment pairs containing the textual information are created on the fly.

This model allows for a very flexible creation of aligned corpora. If the pointers point back into a sparsely annotationed, almost plain, text version of the monolingual corpora then it is easy to work with strings of tokens without the clutter of extraneous elements or attributes that are not always directly useful. On the other hand, if one wants to work with corpora densely annotated with linguistic information such as part of speech tags, lemmatization or phrase mark-up, one only needs to change one line in the alignment corpus to point all references into various versions of the monolingual corpora, which always retain the same unique id-refs irregardless of the extra elements or attributes. The software being used has been developed by the Language Technology Group in Edinburgh and is described in [TFM95]. Gothenburg's implementation is described in more detail in [DR96c] and [Ridings98].

3.1.3 Network access

The database format has the advantage that it is easily integrated with a web server, which permits interactive access across the network. It is currently being totally revamped in order to permit searches on phrases, and automatic suggestions of equivalents in the target language, but an earlier version is still public and can be found through Språkbanken's home page.[4] The new version will eventually replace the present version at the same address.

[4] http://spraakbanken.gu.se

3.1.4 Focus of activities

The initial phases of the work being done in Gothenburg have concentrated on providing auxiliary services to the Faculty of Arts and the translation training program in particular. Early efforts are described in [DR96b]. In addition to interactive tools Gothenburg has also concentrated its efforts on aiding in the boot-strapping of commercial products, that is, supplying them with data. This involves populating translation memories with processed alignment pairs, terminology recognition and streamlining the procedure of collecting and incorporating terminological data for products such as *Multiterm* from TRADOS. The approach to lexical alignment will be described in a forthcoming article in *The International Journal of Corpus Linguistics*.

Research activities are in progress primarily in the form of dissertation topics for graduate students both at the Department of Swedish and the Department of Romance Languages. Pernilla Danielsson, at the Department of Swedish, is working with collocations in two respects: the collocation of lexical units to other lexical units and the association of lexical collocations with the grammatical patterns in their immediate context. The goal is to extend the possibility of isolating translation equivalents from lexical-lexical equivalents to isolating new equivalents based on grammatical patterns around known lexical equivalents. Katarina Mühlenbock, at the same department, is working on transfer rules based on grammatical patterns and Kristina Svensson, at the Department of Romance Languages, is working with a specific semantic field in Swedish (tycka, tänka, tro, anse) and Italian.

3.2 Linköping University

Work on parallel texts in Linköping started in 1993 with the purpose of investigating the use and effects of translation tools of various kinds [AM96]. As part of the project a parallel Swedish/English corpus has been created and a number of tools for alignment and analysis have been developed. The project has had a long-term support from the HSFR/NUTEK Language Technology Programme.

3.2.1 Linköping Translation Corpus

The Linköping Translation Corpus (LÖK) consists of English source texts and their corresponding Swedish target texts which have been aligned on the sentence level. A simple interactive alignment program was developed and used for the purpose [LM94]. The texts are user's manuals for computer programs, novels and a short machine translated dialogue text. The user's guides have been provided by Microsoft Corp. and IBM. The former texts have been translated completely manually, that is without any computational translation support

Sentence-aligned corpus					
	TYPE	DESCR.	SOURCE WORDS	LINKS	METHOD
1.	User's Guide	Microsoft Access	179.631	14.704	Human
2.	User's Guide	Microsoft Excel	149.381	12.589	Human
3.	User's Guide	IBM OS/2	127.499	11.932	Memory
4.	User's Guide	IBM InfoWindows	69.428	7.771	Memory
5.	User's Guide	IBM Client Access	21.321	2.426	Memory
6.	Fiction	Gordimer	197.078	12.254	Human
7.	Fiction	Bellow	66.760	4.209	Human
8.	Dialogue	ATIS	2.179	263	Automatic

Table 1: Linköping Translation Corpus in pure text and Microsoft Access MDB format.

while the latter have been translated with the aid of IBM's Translation Manager (see 2.3.1). The novels were provided by Språkbanken in Gothenburg, while the short machine-translated text was provided by Swedish Institute of Compputer Science (SICS) in Stockholm. A summary of the corpus contents is given in Table 1.

The corpus has a text version and a database version (MS Access). Apart from the sentences of a corresponding pair, a mapping relation between the number of sentences from source to target is recorded for each pair.

A sample of the corpus consisting of some 500 sentence pairs from the Swedish and English versions of each text has been tagged in SGML format for parts-of-speech and lemmatized. These files come in pairs, with a common segmentation into numbered translation units at the sentence level.

One hundred sentence pairs from each of the eight texts have been randomly sampled and tagged for structural and semantic correspondence. This material contains information on the number of translation units, structural translation changes at word and phrase level (such as deletions, additions, convergences, divergences, paraphrases, mode shifts, etc.) as well as changes concerning content and specificity.

3.2.2 DAVE

Several of the programs developed for working with parallel texts have been put together in a software package called DAVE. DAVE stands for "Diagnosis, Alignment and Verification for the Editor". The program is run under Windows 95 or Windows NT and has a graphical interface.

Today DAVE consists of

- A phrase extraction program which retrieves collocations from

a source or target text on the basis on recurrence. This is a new version of FRASSE [MNA94]. A user-defined list of stop words is used to filter the collocations, so that a higher precision is achieved in the search for terms or phrases of given types. Another function of the program is to measure how repetitious a text is, both on the sentence and the phrase level.

- A sentence alignment program that is run in two steps: (i) alignment of paragraphs and (ii) alignment of sentences. This is an interactive, improved version of the LinAlign system used with the LÖK corpus [LM94] Both steps are controlled graphically via a table where the text can be monitored and where potential errors can be fixed easily. The alignment can be run in three modes: automatic, interactive and manual. The result of the alignment is stored as a table which can be edited later and saved as a tab separated text file. Each record in the alignment database consists of the source sentence, target sentence plus information about the number of sentences that are involved in the link.

- A discrepancy program that analyzes a linked translation database (from the alignment program above) as input and returns all inconsistent sentence translations. See [Merkel96] for further details.

- A bilingual concordance program that is used for parallel searches of a translation database. The output from the program is a statistical compilation of the search result and/or all the hits.

3.2.3 Word alignment

The first version of the Linköping Word and phrase Aligner (LWA) has been developed and presented in [AAM98]. LWA takes a sentence-aligned bitext as its input. The output is either (i) the link tokens found in each pair of source and target sentences in the bitext, as in figure 2, and (ii) a bilingual lexicon which includes all link types instantiated by the link tokens from (i).

Work is under way to develop a word and phrase aligner, to be included in DAVE. The system takes a sentence-aligned bitext as input and delivers a set of candidate translations for each pair of corresponding sentences in the bitext. The approach uses co-occurrence statistics as a basis, building on earlier algorithms such as [FC94] and [Mel97b], but implements a number of simple assumptions about the translation process to improve performance. These extra modules are options that the user may ignore, or adjust to his own preferences.

The options that are currently available are:

- The use of function words. These can be included or excluded in the alignment process, and are generically categorized in different subsets (conjunctions, subjunctions, pronouns, etc.) for the

different languages. It is assumed that function words of one category can only be translated with words of a corresponding category in the other language.

- The use of multi-word phrases. These can be included or excluded in the alignment process. When included, phrases are treated on a par with single words.

- The use of inflectional patterns. If some pair $< s, t >$ is selected on statistical grounds, co-occurring pairs consisting of morphological variants of s and t are also selected, even though they have a low frequency and by themselves are not statistically significant.

- The use of a link window. It is possible to restrict the search for target expressions to a window of arbitrary length (in words) around the position in the target sentence that corresponds to the position of the source expression in the source sentence.

- The use of weights based on relative positions in the aligned sentences. Pairs of words having similar relative positions are given higher weights, while pairs of words that are far apart are given low weights. Words outside of the link window are given zero weights.

- The number of iterations. At each iteration the words or phrases that are linked are removed from the list of candidates, thus reducing the size of the bitext and the number of potential candidates for each remaining word or phrase [Mel97b].

- The removal of recurring links. For highly repetitive material, a high number of exact sentence repetitions tend to distort the linking of smaller units contained in these sentences. Therefore it is sometimes necessary to measure the type link only, instead of all the repeated instances of the linked sentences.

All of these options can be used or left out of the alignment process. When phrases and multi-word terminology are used, the source and target texts are first processed with the phrase extracting program mentioned above [MNA94]. The phrases that are retrieved from the texts are used in the program together with a list of general phrasal constructions, thereby handling what can both be the specific terminology for the domain or text type and the general phraseology, often complex subjunctions, adverbials or verbal constructions.

As Swedish words have several morphological variants, a pure string-based alignment approach would fail to find many of the accurate alignments. Thus it would be useful to include a lemmatizer, but wanting this we have designed a simple pattern matcher to handle inflectional variants of regular paradigms. The same mechanism is used for English.

```
I can not accept such a sacrifice from you
Ett sånt offer kan jag inte ta emot av er
```

i <=> jag	(1 <=> 5)
can <=> kan	(2 <=> 4)
not <=> inte	(3 <=> 6)
accept <=> ta emot	(4 <=> 7)
such <=> sånt	(5 <=> 2)
a <=> ett	(6 <=> 1)
sacrifice <=> offer	(7 <=> 3)
from <=> av	(8 <=> 9)
you <=> er	(9 <=> 10)

Figure 2: Sample output from the Linköping word aligner. Numbers in brackets denote sentence positions.

When function words are used, the content words are handled first. After a given set of iterations it is possible to switch to function words and try to align source function words to the class of target function words. Then, in the next iteration, linking of content words and phrases can be resumed.

The output from the alignment is either a list of linked words and phrases sorted alphabetically or by probability. Established links can also be seen graphically with arrows between source and target words and phrases. See figure 2.

In an evaluation performed on one novel (66,693 words) and one program manual (169,779) words, the alignment program showed results of up to 92.5 per cent precision for the novel and 74.93 per cent precision for the manual. When compared to a baseline consisting of the pure statistical algorithm (K-vec), recall was more than tripled at the same time as precision was increased [AAM98].

3.3 Uppsala University

Computational work on parallel text in Uppsala started at the Department of Linguistics in 1993 as a follow-up of the project *Multilingual Support for Translation and Writing*, *Multra*. A concrete result of the project is a prototype of a transfer-based machine translation system, translating from Swedish to English or German (see 3.3.3). The main motivation for initiating work on parallel text was to create a basis for systematically scaling up the linguistic competence of the prototype, thereby turning it into a working translation tool.

In a follow-up project *Multra in action*, *Mia*, a multilingual corpus of technical text with Swedish as the source language was created and a number of tools for processing parallel text were developed. Multra as well as Mia has had long-term support from the HSFR/NUTEK

Language Technology Programme. Mia was conducted in cooperation with Scania CV AB. This cooperation continues in the Scania project (see 3.3.1).

Methodological issues in relation to parallel text are also in focus in the project *Creating and Annotating a Parallel Corpus for the Recognition of Translation Equivalents, Etap* (see 3.3.1). It is part of the Stockholm-Uppsala joint research programme *Translation and Interpreting. A Meeting between Languages and Cultures* (see 3.4).

In the course of the current projects, extensive multilingual materials will have to be administered and continuously updated. For this purpose, quite some effort is devoted to designing and implementing a multilingual lexical database with user-friendly interfaces (see further 3.3.3).

3.3.1 Projects

The Scania project

The basic aim of the Scania project is to create a multilingual machine translation system based on Multra and a controlled version of Swedish, *Scania Swedish*.

As a basis for defining Scania Swedish and for methodological studies and experiments concerning the automatic extraction of translation equivalents, a parallel corpus of maintenance manuals, the *Scania Corpus*, was compiled (see further 3.3.2).

The corpus has been aligned and the tagging of the corpus is in progress. The tagged version of the Swedish part of the corpus will serve as a basis for defining Scania Swedish with regard to sentence grammatical aspects.

A first version of Scania Swedish and a corresponding checker, *ScanCheck*, with a lexicographic tool for updating the vocabulary, *DefLex*, have been developed. They are currently being evaluated on site.

Creating and Annotating a Parallel Corpus for the Recognition of Translation Equivalents, Etap

The basic aim of the *Etap* project is to develop a computerised multilingual corpus that can be used in contrastive lexicographic work, in translation studies, and in methodological studies directed towards the automatic recognition and extraction of translation equivalents from text (see 3.4.

The Etap corpus will comprise Swedish source texts representing different styles and domains with translations into several languages. A basic requirement on the corpus is to have it word class tagged and aligned, primarily sentence by sentence.

So far, three sub-corpora are included in the multilingual Etap corpus, i.e. the Scania Corpus (see 3.3.2), the *Swedish Statement of*

Government Policy Corpus, (see 3.3.2), and the *Immigrant Newspaper Corpus* (see 3.3.2).

The Etap project has a focus on the evaluation and development of methods and techniques for the creation of parallel corpus resources; different kinds of software for format conversion, text processing, alignment, and tagging have been developed. The achievements made in the Etap project are being made available to the other corpus-based projects in the programme. So far, software developed in the Etap project are used in the development of a French corpus(see appendix A) and a Polish corpus(see appendix A.

3.3.2 Corpora

The Scania corpus

The Scania corpus, Scania 9606, is a multilingual collection of truck maintenance manuals of the Swedish company Scania CV AB. The documents are available in 8 European languages: Swedish (source language), Dutch, English, Finnish, French, German, Italian, and Spanish. The original data are stored in Framemaker format. These files were converted to a special SGML format and sentences were aligned (see section 3.3.3).

The text collection is completely parallel. The same information is covered in the same order in every language.

The corpus is extensive. The current size is about 1.6 million words in 80 files for each language part. The following table shows size information to each language part.

language	files	words	bytes
Swedish	80	172259	7792597
Dutch	80	216424	8072128
English	80	220827	7886082
Finnish	80	148348	7833990
French	80	244239	8156457
German	80	186293	8004331
Italian	80	228631	8127121
Spanish	80	250730	8090916
total	640	1667751	63963622

The text documents contain typical technical descriptions of maintenance tasks. The text is written in a brief manner, and information is usually given in short clauses or phrases. Sentence structure is usually simple. Complex sentences with subordinate clauses are rare.

Special structures such as lists, tables, and figures are frequent. Many paragraphs and sections are short and there is a large number of headers. The headers represent phrases rather than sentences.

Many one-word units appear as independent segments, for instance in table cells. Tables, lists, and labels of various kinds are also parallel structures in the different languages. They represent a good starting point for finding corresponding parts in the multilingual corpus.

Another characteristic feature of the corpus is the large number of technical words and abbreviations. Technical names are usually similar in different languages and therefore they can often be identified as cognates. Some of them occur as identical copies in different languages.

Due to these characteristics the Scania corpus provides a promising resource for multilingual information retrieval.

The Swedish Statement of Government Policy Corpus

The Swedish Statement of Government Policy Corpus, Regeringsförklaringen (RF 9607), is a collection of Government Statements made in 1988 (Swedish, English, German, French), 1994 (Swedish), 1995 (Swedish), 1996 (Swedish, English, German, French, and Spanish). The total size of the corpus amounts to 26,709 tokens. The following table shows size information about each language.

language	files	words	bytes
Swedish	4	10419	140924
English	2	4492	63522
French	2	5221	67769
German	2	4259	67650
Spanish	1	2318	30930
total	11	26709	370795

The corpus is available at http://strindberg.ling.uu.se/ corpora/rf. A demo of a software for searching the corpus can be found at the same location.

The Immigrant Newspaper Corpus

Work on the collection of material from the Immigrant Newspaper, 'Invandrartidningen' is in progress. The newspaper is published in eight language versions, i.e. simplified Swedish, English, Finnish, Polish, Serbo-Croation, Spanish, Arabian, and Persian. The translations are based on a Swedish original text that is not published. Instead, a simplified version of the original text 'På lätt svenska' is published. In 1997, seven issues of the newspaper in seven language versions (original source version, simplified version in Swedish, English, Finnish, Polish, Serbo-Croation, and Spanish) were delivered via ftp to the department. This material will provide roughly 200,000 current words (5,000 current words pro issue and language version). The Swedish

original text of the first 20 issues of 1997 with translations into the same target languages were delivered as hard copy. This text will be fed into the computer via optical character recognition. It will add some 650,000 current words to the corpus. No Arabian and Persian translations are available in machine-readable form, and for the time being these language versions will be left outside the corpus. The total size of the corpus will amount to 850,000 tokens.

The extra work needed for creating a computerized corpus of the multilingual material from the Immigrant Newspaper seems to be worthwhile due to the variety of the languages that it comprises. It will provide an interesting research material for contrastive investigations and multilingual lexicography. Currently, parts of the material are being used as a testbed for research on measures for string similarity and the use of cognates in word alignment and the extraction of lexical translation equivalents ([Bor98]). The translation equivalents will serve as a basis for the extraction of lexical, in particular, morphological information about the target languages. However, prior to this, the following steps have to be taken:

- conversion of the Immigrant Newspaper PageMakerfiles into HTML
- scanning and proof-reading of the hardcopy source and target versions
- conversion of scanned versions into Uppsala-SGML
- sentence alignment Swedish - English, Swedish - Finnish, Swedish - Polish, Swedish - Serbo-Croatian, Swedish - Spanish

3.3.3 Methods and Tools

Machine translation - MULTRA

Multra is a modular transfer-based machine translation system [ASH95]. It works in following four steps.

Analysis: An analyser generates internal representations of grammatical structures in the source language in terms of attribute-value structures; the grammatical structures express grammatical function and constituency

Preference: A preference component orders the grammatical structures in a linguistically preferred order and presents them to the transfer component

Transfer: A transfer component applies transfer rules to the analysis structures and generates target language structures; transfer is realised as unification of attribute-value structures; a special formalism for the expression of transfer rules has been defined; there is no formal difference between lexical and structural transfer rules; lexical units can be transfered out of context

or in context; in the latter case, the transfer rules are defined to cover an appropriate section of the analysis structure

Generation: A generation component generates target language expressions from the grammatical structure; a special formalism for the expression of generation rules has been defined;

Converting to SGML

The original files of the Scania truck manuals were written in Framemaker. The internal format of the Framemaker files includes page style formats, text styles, pictures and other kinds of typographic information. Their usage is limited to platforms with appropriate software. In order to overcome this problem, the documents were converted to SGML [San96b].

As document style definition, a subset of the encoding scheme defined by the Text Encoding Initiative Guidelines [SMB94] *TEI Lite*, was chosen. For the conversion from the Framemaker format to TEI Lite tagged files, a conversion tool written by Ken Harward was used. This software *MifMucker* [Har94] is a collection of filters which includes a tool for the conversion of documents from the Maker Interchange Format (MIF) - the interchange format of Framemaker - to HTML. This software (Perl scripts) was modified to produce TEI-Lite-conformed documents from the Framemaker source files.

In addition to this conversion, sentence boundary tags were added.

The Swedish Statement of Government Policy corpus is based on plain ASCII text files. These texts were converted to SGML with the same data type definition that was used for encoding the Scania corpus files (Uppsala SGML - a subset of TEI Lite).

The Immigrant Newspaper corpus is based on Pagemaker files and paper versions. Pagemaker provides a conversion to HTML which can then be used for the conversion to *Uppsala SGML* format. The printed versions have to be scanned and processed via OCR software. After manual inspection, and, maybe, correction the resulting ASCII files will be converted to Uppsala SGML.

Sentence Alignment

For multilingual sentence alignment the method proposed by Gale and Church [GC93] was chosen (see 2.1). When applied to the Scania corpus, it provided quite good results, i.e. for Swedish/Finnish (about 97% correct alignments) , and for the other language pairs somewhat worse (about 90-91% correct alignments). Two paragraph errors were detected, and when they had been removed, the success rate exceeded 99% (see further [San96a]).

The same algorithm was applied to the Swedish Statement of Government Policy Corpus.

Tagging

Several methods for word alignment and the identification of translation equivalents are based on tagged corpora. Consequently, taggers belong to the tool kit required by the parallel corpus worker. In the Etap-project training of Brill taggers ([Bri92]; [Bri95]; [Bri97]) for general Swedish ([Prü97]), for the technical Swedish of the Scania Corpus ([Ekl98]), and for French ([Nyl97a]) is in progress.

A first version of the general Swedish Brill tagger has been integrated into the Gate platform ([Ols97]) from where it is generally available for research purposes.

Extraction of Translation Equivalents - LexEx

The focus of another study, LexEx, based on the aligned parallel Scania corpus was the extraction of translation equivalents from bilingual alignments [Tie97]. Within this project currently used methods and approaches were applied to the Swedish/English and Swedish/German alignments of the corpus. The approaches include:

Extraction by size reduction: This approach focuses on short aligned structures. Using a basic dictionary alignments are reduced in size in an iterative process and remaining data are analysed.

String similarity evaluation: Simple string comparing algorithms based on character matching are used to value the similarity between word pairs. Evaluations by threshold filtering and score combination techniques produced a set of cognate pairs with reasonable precisions for the considered language pairs.

Statistical evaluation: This approach uses statistical measures based on frequency counts to identify pairs with a high co-occurrence ratio. The Dice coefficient [SMH96] was used to value word pairs which were compiled from bilingual alignments.

Evaluations of low frequent words: In this approach the assumption was used that low frequent text units are translated into low frequent text units in other languages. For this purpose, high and medium frequent words were removed from the alignments and remainig data were analysed to find corresponding low frequent translation equivalents.

Detection of compounds: Word mappings based on similarity measures were used to detect translation equivalents with different usage of compound structures. This approach focuses on extracting pairs with compositional compounds in one language and non-compositional compounds in the other language.

Group by part of speech: The Swedish/English part of the Scania corpus was partially tagged with part-of-speech tags using the morphological analyses generated by the Uppsala Chart

Parser ([ASH81]) and using a small tagged English dictionary ([Kei]). Now, all words with the same part-of-speech tag from one sentence alignment were aligned to each other. The resulting part-of-speech alignments were analysed to extract corresponding word pairs.

The final part of the project described above was concentrated on automatic filtering processes of the produced translation pairs. For this purpose several approaches based on statistical and empirical evaluations were used to remove obviously wrong pairs.

Lexical Database

In current work a flexible structure for a linguistic database is under development. The database will include lexical data in form of morphological, syntactic, and semantic information about lexical terms as well as corpus data. Current investigations consider relational approaches for the storage of lexical data. Corpus data are stored in SGML format (conforming TEI encoding standards) and will be linked to the lexical database via indexing. The main purpose for compiling a global database structure is flexible and fast access to the complete set of collected linguistic data. A set of tools are developed to manage the lexical database. These tools include functionality for creating database structures, adding data from different ressources (e.g. UCP parses, stem lexica, SGML encoded corpora), and searching the database.

Interactive User Interfaces

XLexEx : A graphical user interface (XLexEx) was developed to combine lexicon extraction methods developed and implemented within the LexEx study. This tool includes several modules to setup configuration data, call external programs and filter, and to examine results. In the current stage the tool provides modules for preparing bilingual alignments for further processing, extracting arbitrary alignments from the corpus, applying several extraction methods, converting files to different formats, viewing and counting input files and resulting files, merging results, configuring and applying automatic dictionary filter, and looking for concordances in the aligned corpus. The system is window oriented and highly configurable. Additional modules are in preparation.

QLexDB : QLexDB represents a graphical user interface for querying and updating lexical data in the current version of the relational lexical database. In the current implementation, morphological and syntactic information can be searched in the database and morphological data can be modified and added. Furthermore, the tool provides

queries for concordances and alignments in the Scania corpus. Results from database queries can be saved in ASCII format for further processing.

WWW-Interface for Corpus and Lexical Data : A set of HTML encoded pages are written using CGI scripts to access the current lexical database. The tools include forms for searching the UCP stem lexicon, for searching morphological and syntactic data in the lexical database, for finding concordances and alignments in available SGML encoded corpora, and to query corpus indices.

3.4 Other Swedish projects on parallel corpora

Lund University and the Scandinavian Project of Contrastive Corpus Studies

There is an ongoing Scandinavian project involving four partners in Norway, Finland, Denmark and Sweden. Swedish is represented by the Department of English at Lund University with their "Text-based Contrastive Studies in English" [AAJ96]. Norway is represented by those who were already involved with the English-Norwegian Parallel Corpus, ENPC and is described in [JH94]. Finland's involvement is by way of the Finnish-English Contrastive Corpus Studies (FECCS) project at the Department of English, University of Jyväskylä, Finland. We have no information about the Danish project. All four corpora will be built up according to the same structure. Each corpus consists of two parts: one parallel corpus of original texts together with their translations, and one comparable corpus consisting of original texts in both languages. The corpora are to be used in contrastive studies between the Scandinavian languages and English.

The parallel corpus between Swedish and English is planned to consist of 1,600,000 words in different samples (each sample 10,000–15,000 words) from both directions. The corpus will become available as soon as all the copyright restrictions are resolved.

Stockholm-Uppsala research program on translation and interpreting.

Translation studies based on parallel corpora are carried out at several departments in the framework of the joint Stockholm-Uppsala research program on translation and interpreting. *Translation and Interpreting. A Meeting between Languages and Cultures*, financed by the Bank of Sweden Tercentenary Foundation 'Riksbankens Jubileumsfond' (see further [SU95]). Corpus-based contrastive studies is one of three dimensions along which research is directed in the programme. It implies contrastive studies of original text and its translations. What are the characteristics of the target language texts? How are they related to the source language texts? To what degree have they been adapted to the norms of the target language? These

are some of the questions asked in relation to the study of this dimension. The text material that is examined consists of both literary texts and and LSP[5]. Research in machine translation is an integrated part of this aspect of the programme, and one of the aims of this research is to contribute to the development of theories about complex phenomena such as referential expressions, metaphors, discourse markers, and phraseology [Sva97].

A great variety of languages, styles and domains, Swedish both as a source and as a target language, written and oral text, and the many different research issues involved give a special profile to the corpus studies of this research programme. In the Appendix, the corpus-directed projects of the programme are listed.

Swedish MULTEXT at University of Umeå

The Department of Linguistics at the University of Umeå was an associated partner of the European LRE project "Multilingual Text Tools and Corpora" (MULTEXT, LRE 62 – 050). The aims of the MULTEXT project is to develop tools and to prepare comparable, parallel and speech corpora for seven EU languages: English, French, German, Italian, Spanish, Dutch and Swedish. The Swedish participation in MULTEXT is financed by NUTEK and work on the project was carried out during 1994/95 and 1995/96. The final deliverables for Swedish consist of the Swedish part of a comparable corpus of financial newspaper texts, a speech corpus, and lexical lists.

4 Summary and outlook

It can be said with certainty that parallel corpora are very useful sources for a range of subfields in linguistics and for multilingual language technology. There are several tools available that support alignment on the sentence level, and for searching them, once they are aligned. This does not mean, however, that everything is done. Moreover, there are a number of problems that prevent a rapid exploitation of parallel corpora.

4.1 Acquisition of parallel corpora

First of all, parallel texts are not easily available, since most documents are not published in parallel. Moreover, distribution of texts is subject to copyright restrictions and the copyright holder for a source text is not necessarily the same as the copyright holder for the translation. The creation of large, balanced parallel multi-lingual corpora, which would be so valuable both to research communities and technical development is thus still a goal, which requires the cooperation of many parties.

[5] Language for Special Purposes

Text capture also often requires substantial efforts. If you get the text on paper, you will either have to enter it manually into the computer or use a scanner with character recognition capabilities. Both methods may introduce errors, that you might want to correct and this will require further effort. If the texts are delivered in a word processor format other than that your project is using, you need to get rid of the formatting instructions. Conversion to text can of course often be done automatically, but this means that you lose information about text structure and block types that would be required, or at least useful, for the later processing of the texts. Thus, much more effort usually has to be spent on text capture and cleaning up formatting commands than on alignment. These problems are not easy to overcome and have to be taken into account in any project on parallel texts.

4.2 Encoding Standards

Another important issue concerns annotations and encoding standards. While the TEI (Text Encoding Initiative [SMB94] provide several suggestions as to the representation of links, there remains a lot to be done as regards standards. Such work is now being undertaken by both the TEI and EAGLES.

An important problem, which has been the concern of the PEDANT project since early 1996, is the search for efficient and tractable ways to encode multilingual corpora in order to retain their value as straight monolingual corpora yet record the additional information such as alignment between segments, grammatical information etc. A detailed solution is proposed in Ridings [DR96c]. See also 3.1.2 for some discussion.

New general standards are emerging, however, that address such problems, namely The Extensible Markup Language, XML. This is a simplified form of SGML that is being defined by the WWW Consortium [BSM96] and its connection with the web is bound to open new possibilities for presenting material on the network, be it Internet or intranet.

In addition to the more general standards, there is an effort to define a translation database exchange format, which is closely followed in the PEDANT project.

4.3 Data extraction issues

Work on aligning parallel texts at the word and phrase levels have just begun, but existing results in this area are quite promising. Still, there will probably be much improvement as more experienced is gained. Existing systems exploit a number of different knowledge sources and heuristics, but so far these have not been explored systematically. As phrase extraction tools, taggers, lemmatizers and partial parsers are developed and applied to bitexts, current systems

will show higher performance. Close analyses of alignment results will probably also result in a better understanding of the translation process and the type of correspondences that current systems miss out.

There exist a number of even more challenging tasks that are beginning to be investigated. These concern the generation of complex translation data such as contrastive valency data and transfer rules for machine translation systems. Another line of research seeks to develop methods for adapting an existing general knowledge source such as a bilingual or multilingual dictionary to the domain and phraseology of a given text type.

4.4 Evaluation

So far there are no generally agreed methods for evaluating software for processing parallel texts. Most researchers report results for the parallel texts that have been available to them, which is usually not the same texts that other researchers have available, and also the measures used tend to vary from one group to another. This is obviously something that must change in the future. A step in this direction has been taken by the ARCADE project [6] which attempts to develop generally agreed measures both for sentence alignment and word alignment, and organises "competitions" to try them out in practice. ARCADE is currently restricted to French and English.

In the case of sentence alignment comparison with a gold standard, a completely aligned and checked bitext, seems feasible, and measures of success based on precision seem relatively easy to agree upon. However, one must bear in mind that one error may cause complete failure for whole sections of a text, and conversely, that a result that looks disastrous on the basis of counting hits, may be rectified by correcting a single error. For this reason, measures that take into account the human effort to correct an automatically generated sentence alignment are also valid.

Also for word alignment comparison with human performance has been suggested, although in this case much more work is needed to create the basis for comparison. There are projects, however, that have created such annotations, the ARCADE project, mentioned above and the Blinker project at University of Pennsylvania[7].

A significant complication is that it is not always clear what should count as corresponding items, even for humans. Consider the following examples:

Eng. *He was dreaming about going to the States.*
Swed. *Han drömde om att åka till USA.*

In this case we may take either *was dreaming* or *dreaming* as corresponding to Swedish *drömde*. How we choose depends on whether we

[6] http://www.lpl.univ-aix.fr/projects/arcade/index-en.html
[7] http://www.cis.upenn.edu/~melamed

are primarily interested in textual or lexical correspondences. Such analysis problems are quite common; in this example there is a similar uncertainty about the correspondence *going - (att) åka*.

Another common case of dispute is the case of collocations. If, say, *New York Times* occurs in both the original and the translation, do we have one alignment between two identical items consisting of three words, or do we have three alignments between three one-word items, or perhaps two alignments, one for the collocation *New York* and one for the single word *Times*?

A simpler use of an existing translation has been proposed by Dan Melamed [Mel95]. His suggestion is that a small part of the corpus is held out for test purposes. The lexicon generated from the remaining corpus is applied to the source half of test corpus in order to find out what percentage of the words in the other half the lexicon can account for. An advantage with this method is that no human annotation is necessary.

4.5 Deployment

While parallel corpora are beginning to be exploited in many areas, there are still many open questions about the proper ways of doing it, i.e. how to fit data generation tools and bilingual concordances into the environments and work processes that translators, terminologists, lexicographers and language teachers are using. While good examples exist e.g. in the case of terminology (cf. 2.3) there remains to be seen how these tools can be applied in other areas.

Also for automatic systems there is some way to go before a parallel corpus can be used as the primary input. Current methods for statistical machine translation are not sophisticated enough to derive good translation models even from very large parallel corpora. Thus, also in the area of machine translation the primary use to date of these methods seem to be for manually updating and extending subject dictionaries.

4.6 The PLUG project

A number of research issues have been mentioned in passing in the previous sections of this chapter and it is to be hoped that they will be tackled in due course. Some of these, which appear to us to be of immediate importance and within reach, will be worked on in the project *Parallel Corpora in Linköping, Uppsala and Göteborg* (PLUG). The goals of this project are:

- Collection and annotation of a parallel sentence-aligned corpus with Swedish-English, Swedish-German, Swedish-French and Swedish-Italian texts of different genres.
- Design and trial of evaluation methods for word and phrase alignment systems.

- Design and implementation of new systems for word and phrase alignment and for generation of construction data.
- Design and implementation of a lexical database with links to the project corpus
- Formalization of lexical and construction data in accordance with the demands of a transfer-based MT-system

References

[AAJ96] Karin Aijmer, Bengt Altenberg and Mats Johansson Text-based contrastive studies in English. Presentation of a project. In *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994.* Lund University Press.

[AAM98] Lars Ahrenberg, Mikael Andersson and Magnus Merkel A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. In *Proceedings of COLING-ACL'98, August 10-14, 1998, Université de Montréal, Montréal, Canada, pp. 29-35.*

[AH96] Ingrid Almqvist and Anna Sågvall Hein. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications*, KU Leuven, Belgium, 1996.

[AM96] Lars Ahrenberg and Magnus Merkel On translation corpora and translation support tools: A project report. In *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994.* Lund University Press.

[ASH81] Anna Sågvall Hein. An Overview of the Uppsala Chart Parser Version I (UCP-1). Technical report, Department of Linguistics, University of Uppsala, 1981.

[ASH93] Anna Sågvall Hein. Multilingual Support for Translation and Writing. MULTRA. Technical report, HSFR/NUTEK Language Technology Research Program, 1993.

[ASH95] Anna Sågvall Hein. Preference Mechanisms of the Multra Machine Translation System. In *Hall Partee, B. & P. Sgall (eds.) Menaing and Discourse. Festschrift für Eva Haijcova.*, J. Benjamin's Publishing Co, 1995.

[ASH97a] Anna Sågvall Hein. Language Control and Machine Translation. In *Proceedings of the 7th Conference of Theoretical and Methodological Issues in Machine Translatio*, Santa Fe/New Mexico, 1997.

[ASH97b] Anna Sågvall Hein. The Morphological Description of Sve.Ucp. Technical report, Department of Linguistics, University of Uppsala, 1997.

[BSM96] T. Bray and C. M. Sperberg-McQueen. Extensible markup language (xml) version 1.0. Working Draft WD-xml-961114, World Wide Web Consortium, 1996. Available at `http://www.w3.org/pub/TR/WD-xml-link`.

[BR95] Patrice Bonhomme and Laurent Romary The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purposes. Language Engineering 95, Montpellier June 26-30, 1995.

[Bor98] Lars Borin. Linguistics isn't always the answer: Word comparison in computational linguistics. accepted for the 11th Nordic Conference on Compuational Linguistics NODALI98. Department of Linguistics, University of Uppsala, 1998

[BCDD90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John. D. Lafferty and Robert L. Mercer. *A Statistcal Approach to Machine Translation.* Computational Linguistics, 16(2), 79-85, 1993.

[BDDM93] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. *The Mathematics of Statistcal Machine Translation: Parameter Estimation.* Computational Linguistics, 1993.

[Bri92] Eric Brill. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento/Italy, 1992. Association for Computational Linguistics.

[Bri95] Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics, December*, 1995.

[Bri97] Eric Brill. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. Kluwer Academic Press, 1997.

[CGHH91] Kenneth W. Church, William Gale, Patrick Hanks, and Donald Hindle. Using Statistics in Lexical Analysis. In Uri Žernik, editor, *Lexical Acquistition: Using on-line resources to build a lexicon.* Lawrence Erlbaum, 1991.

[CH93] Kenneth W. Church and J. Helfman. Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code. The Journal of Computational and Graphical Statistics, 1993.

[Chu93] Kenneth W. Church. Char_align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, ACL.* Association for Computational Linguistics, 1993.

[CWR90] Roberta Catizone, Graham Russell and Susan Warwick. Deriving translation data from bilingual text. In *Proceedings of the First International Lexical Acquisition Workshop*, IJCAI, 1989.

[DC94] Ido Dagan and Kenneth W. Church. Termight: Identifying and Translating Technical Terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart/Germany, 1994. Association for Computational Linguistics.

[DCG93] Ido Dagan, Kenneth W. Church, and William Gale. Robust Bilingual Word Alignment for Machine-Aided Translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Respectives*, Columbus/Ohio, 1993. Association for Computational Linguistics.

[Dic45] Lee R. Dice. Measures of the Amount of Ecologic Associations between Species. *Journal of Ecology, 26*, 1945.

[DR97] Pernilla Danielsson and Daniel Ridings. Practical presentation of a vanilla aligner. Technical report, Språkbanken, Institutionen för svenska språket, Göteborgs universitet, 1997.

[DR96c] Pernilla Danielsson and Daniel Ridings. Pedant: Parallel texts in göteborg. Research reports from the Department of Swedish, Göteborg University GU-ISS-96-2, Språkdata, 1996.

[DR96b] Pernilla Danielsson and Daniel Ridings. Corpus and Terminology: Software for the Translation Program at Göteborgs universitet or Getting students to do the work. Proceedings from Teaching and Language Corpora, Lancaster, England, 1996.

[DR96c] Pernilla Danielsson and Daniel Ridings. Annotating Parallel Texts with the NSL library, Technical report GU-ISS-96-?, Språkdata, Department of Swedish, Göteborg University, 1996.

[Ej+92] E. Ejerhed et al. The lingusitic annotation system of the Stockholm-Umeå Corpus Project. University of Umeå, 1992.

[Ekl98] Maria Eklund. The Brill Tag Set of the Scania Corpus Technical report, Department of Linguistics, University of Uppsala, 1998.

[FM94] Pascale Fung and Kathleen R. McKeown. Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In *Proceedings of the 1st Conference of the AMTA*, Columbia/Maryland, 1994. Association for Machine Translation in the Americas.

[FC94] Pascale Fung and Kenneth W. Church. K-vec: A new approach for aligning parallel texts. In*Proceedings from the 15th International Conference on Computational Linguistics (Coling-94)*, pages 1096–1102. Kyoto, 1994.

[GC93] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19(1)*, 1993.

[Har94] Ken Harward. MifMucker, Application for manipulating Frame documents and books. Availabe from http://www.oac.uci.edu/indiv/ehood/mifmucker.doc.html, 1994.

[HAS97b] Anna Sågvall Hein and Ingrid Almqvist and Per Starbäck. ScaniaSwedish - a Basis for Multilingual Machine Translation. In , London, UK, 1997.

[Hofl95] Knut Hofland. A program for aligning English and Norwegian sentences. *Proceedings of the ACH/ALLC Conference.* Santa Barbara, USA, July 1995.

[Hutchins:86] W. J. Hutchins. Machine Translation: Past, Present, Future Ellis Horwood, Chichester, 1986

[IMS97] The IMS - Corpus Toolbox. http://www2.ims.uni-stuttgart.de/oli/CorpusToolbox/CorpusToolbox.h tml, 1997

[JH94] Stig Johansson and Knut Hofland. Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie, and

P. Scheider, editors, *Creating and Using English Language Corpora*, pages 25–37. Editions Rodopi, Zrich, 1994.

[Kei] Dave Keiras. English lexicon. Software availabe from University of Michigan.

[KT96] Judith L. Klavans and Evelyne Tzoukermann. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation*, 1996.

[Kup93] Julian M. Kupiec. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. In *Proceedings of the 31st Annual Meeting of the ACL*, Columbus/Ohio, 1993. Association for Computational Linguistics.

[LP96] Josette Lecomte and Patrick Paroubek. Le Catègoriseur d'Eric Brill (U. Penn) Mise en Œuvre de la Version Entraînèe pour l'INaLF. INaLF - Nancy, 1996.

[Leech93] Geoffrey Leech. The state of art in corpus linguistics. In K. Aijmer and B. Altenberg, eds. *English Corpus Linguistics. Studies in honour of Jan Svartik.* Longman, London, 1993.

[LM94] Arne Larsson and Magnus Merkel. Semiotics at Work: Technical Communication and Translation in a Multilingual Corporate Environment In NODALIDA '93 (Proceedings of 9:e Nordiska Datalingvistikdagarna', Stockholm 3-5 June 1993, Department of Linguistics, Stockholm, 1994.

[LTN97] LT NSL - The Normalized SGML Library. http://www.ltg.ed.ac.uk/corpora/nsldoc/nsldoc.html, 1997.

[Mel95] I. Dan Melamed. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*, Boston/Massachusetts, 1995.

[Mel96] I. Dan Melamed. A Geometric Approach to Mapping Bitext Correspondence. In *Conference on Empirical Methods in Natural Language Processing*, Philadelphia/USA, 1996.

[Mel97b] I. Dan Melamed. A word-to-word model of translational equivalence. In *35th Conference of the Association for Computational Linguistics (ACL'97)*, pages 490–497, Madrid, 1997.

[Merkel96] Magnus Merkel. Checking translations for inconsistency - a tool for the editor. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96)*, pages 157–167. Montreal, 1996.

[MNA94] Magnus Merkel, Bernt Nilsson, and Lars Ahrenberg. A phrase-retrieval system based on recurrence. In *Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2)*, pages 99–108. Kyoto, 1994.

[MUL97] Multext tools. http://www.lpl.univ-aix.fr/projects/-multext/MUL7.html, 1997.

[MWSS97] Tony McEnery, Andrew Wilson, Fernando Sánchez-León and Amalio Nieto-Serrano. Multilingual Resources for the European Languages: Contributions of the CRATER Project. *Literary & Linguistic Computing*, Vol. 12 No. 4, November 1997.

[Nyl97a] Stina Nylander. Utveckling av en Fransk Brilltaggare Technical report, Department of Linguistics, University of Uppsala, 1997.

[Nyl97b] Stina Nylander. The Uppsala Brill Tagger for French Technical report, Department of Linguistics, University of Uppsala, 1997.

[Ols97] Fredrik Olson. *Tagging and Morphological Processing in the SVENSK System.* Master thesis, University of Uppsala, Department of Linguistics, 1997.

[Öst87] Andersson A. Östling. L'identification automatique des lexèmes du français contemporain. Uppsala, 1987.

[Prü97] Klas Prütz. Sammanställning av en träningskorpus på svenska för träning av ett automatiskt ordklasstaggninssystem. Technical report, Department of Linguistics, University of Uppsala, 1997.

[Ridings98] Daniel Ridings. PEDANT: Parallel Texts in Göteborg, forthcoming (Sept. 1998) Lexikos, Vol. 8, 1998.

[Sag+90] Anna Sågvall Hein and Eva Wikholm and Annette Östling. Phrases in the Core Vocabulary - A Contrastive Study of the Statements of Government Policy 1988. Uppsala University, Department of Linguistics, UCDL-R-90-1, repro HSC, 1990.

[San96a] Erik F. Tjong Kim Sang. Aligning the Scania Corpus. Technical report, Department of Linguistics, University of Uppsala, 1996. Available at http://stp.ling.uu.se/erikt/papers/

[San96b] Erik F. Tjong Kim Sang. Converting the Scania Framemaker Documents to TEI SGML. Technical report, Department of Linguistics, University of Uppsala, 1996. Available at http://stp.ling.uu.se/erikt/papers/

[SFI92] Michael Simard, George F. Foster, and Pierre Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal/Canada, 1992.

[SMB94] C.M. Sperberg-McQueen and Lou Burnard. Guidelines for Electronic text Encoding and Interchange. Availabe from http://etext.virginia.edu/TEI.html, 1994.

[Sma93] Frank Smadja. *Retrieving Collocations from Text: XTRACT.* Computational Linguistics, 1993.

[SMH96] Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translation Collocations for Bilingual Lexicons: A Statistical Approach. In *Association for Computational Linguistics.* Association for Computational Linguistics, 1996.

[SN90] S. Sato and Makato Nagao. Towards Memory Based Translation. In *Proceedings from the 13th International Conference on Computational Linguistics (COLING-90)*, pp. 247-252, Helsinki, 1990.

[SOB86] A dictionary of Swedish. Stockholm, 1986.

[SU95] Språkvetenskapliga sektionerna vid universiteten i Stockholm och Uppsala. RJ 1995-08-18. Översättning och tolkning som språk- och kulturmöte. Språkvetenskapligt forskningsprogram. Stockholms universitet. Uppsala universitet. 1995.

[Sva97] Brynja Svane. Creating and Annotating a Parallel Corpus for the Recognition of Translation Equivalents. In *Translation and Interpretion.*, Repro HSC, Uppsala Universitet, 1997.

[TFM95] Henry Thompson, Steve Finch and David McKelvie. The Normalised SGML Library (NSL). The Language Technology Group, LRE Project 62-050 (Multext), November, 1995.

[Tie97] Jörg Tiedemann. *Automatical Lexicon Extraction from Aligned Bilingual Corpora.* Diploma thesis, University 'Otto-von-Guericke', Magdeburg, Department of Computer Science, 1997.

[TIP97] TIPSTER Text Program http://www.tipster.org/, 1997.

[vdE93] Pim van der Eijk. Automating the Acquisition of Bilingual Terminology. In *Proceedings of the 6th Conference of the European Chapter of the ACL*, Utrecht/The Netherlands, 1993. Association for Computational Linguistics.

[Wik+93] Eva Wikholm and Annette Östling and Ingrid Maier. A Multilingual Dictionary of Functional Core Phrases with Prepositions. A MULTRA-Report, Uppsala University, Department of Linguistics, repro HSC, 1993.

[WX94] Dekai Wu and Xuanyuin Xia. Learning an English-Chinese Lexicon from a Parallel Corpus. In *Proceedings of the 1st Conference of the AMTA*, Columbia/Maryland, 1994. Association for Machine Translation in the Americas.

A Appendix

Contrastive, corpus-based projects in the research programme Translating and Interpreting, a Meeting between Cultures.

- French Cultural Images in a Swedish Context, the Department of French and Italian, SU, directed by professor Brynja Svane. Corpus: *a French-Swedish corpus of literary text and LSP text*
- Literary Translation as Cultural Transfer - A Descriptive Analysis of Literary Translation Strategies, the Department of Nordic Languages, SU, directed by professor Staffan Hellberg. Corpus: *an English-Swedish corpus of literary text*
- Gender and Sex in German and Swedish - A Contrastive Analysis, the Department of German Language, SU, directed by associate professor Gunnar Magnusson. Corpus: *German-Swedish newspaper corpus*
- The Translation of Oral Texts from Indigeous Languages, the Department of Linguistics, UU, directed by associate professor Anju Saxena. Corpus: *a bilingual corpus of Kinnauri text and its translation into English*
- The art of translating from French to Swedish, the Department of Romance Languages, UU, directed by professor Kerstin Jonasson. Corpus: *a bilingual text corpus of French source texts and their Swedish translations*
- Translation - a Dimension in the History of Swedish, the Department of Nordic Languages, UU, directed by associate professor Lars Wollin. Corpus: *a corpus of Swedish translations from English, French, Spanish, German, Italian and Danish from the 20th century*
- The Perception of Polish Literary Texts through Swedish Translations, the Department of Slavic Languages, UU, directed by professor Sven Gustavsson. Corpus: *a bilingual corpus of Polish source texts and their Swedish translations*
- Creating and annotating a parallel corpus for the recognition of translation equivalents, Etap, see further above 3.3.1