Published in "Languages in Contrast. Papers from a Symposium on Text-bsed cross-linguistic studies. Eds. Aijmer, K., Altenberg, B. & Johansson, M. Lund University Press, 1996.

On Translation Corpora and Translation Support Tools: A Project Report

Lars Ahrenberg and Magnus Merkel
Department of Computer & Information Science
Linköping University

For more than forty years machine translation has been an active area of research and development. Several variants of MT systems have been proposed e.g., interactive systems, batch-mode systems requiring a large amount of human post-editing and systems that generate text in several languages from a common structured set of data rather than a source text. Although there are some successful systems and several products on the market, it is still the case that most translators and translation companies prefer not to use MT-systems. In fact, it has been estimated “that the current share of anything that could be called MT, pure or mixed, is well below 1% of the total translation market” (Isabelle et al, 1993).

This situation is likely to change, however. The disappointment over the performance of automatic systems has caused a change of direction in system development, putting the focus on systems that will support the translators rather than replace them. The ever increasing processing capacity of PCs and workstations will also enable translators and companies to purchase powerful tools as add-ons to their ordinary word processing programs. Furthermore, as we show below, already very simple tools can speed up the translation process for certain text types by 30-40%.

The experiences of the past forty years of MT research can be summarised as follows: (i) there is a negative correlation between the size and complexity of the intended domain and the extent to which the translation/generation process can be automated; fully automatic document generation and translation has been successful only with small sublanguages such as the language of weather or stock market reports (Kittredge et al. 1986); (ii) hence, for most domains it is necessary to consider the human involvement in the process, and, especially in large-scale translation efforts, the overall organisation and control of the process; (iii) there is a great need, and today a great opportunity, to customise a system on the basis of relevant data and to develop tools that can access and extract the data from relevant sources.

The failures of the past have also lead to the consideration of new methods. An idea that has attracted a great deal of interest in recent years is the memory-based system. The basic idea of this kind of system is that of reusing earlier translations. These are stored in a database - *the translation memory* - where sentences of the source text have been linked to corresponding sentences of the target text. The result is a structure which is sometimes referred to as a *bi-text* (Isabelle et. al. 1993).

The translation memory may be used in different ways in the translation process. Prior to translation it may be used to identify units of a new text that can be found in the memory, during translation to answer specific queries from the translator and after translation to check consistency with previous translations.

Translation memories may also be used as part of automatic systems. Such systems are called example-based and usually store the translation units in analysed form (Nagao, 1984; Sato & Nagao, 1990). The assumed advantage of an example-based system over a more conventional rule-based MT-system is that it can handle cross-linguistic differences at a finer level of detail. Moreover, it is designed to draw analogies from the examples in its memory and combine the results of partial matches into adequate target language sentences.

The difference between a translation memory as a support tool for the translator and a full-fledged example-based system is thus basically a difference in who has the prime responsibility for drawing analogies and structure the target text during translation. In either case the translation memory is a key knowledge source, which provides access to the efforts and results of previous work. A translation memory is a particular case of a parallell corpus and could therefore provide interesting data for cross-linguistic studies and the study of language use in translations.

In this paper we report on some preliminary results of a project at Linköping university studying how information in existing translations and source texts can be exploited in different phases of a translation project.[1] This is one aspect that can contribute to better translation support systems, but there are other important aspects, such as the user interface and the overall system design, but these fall outside the scope of this particular project.

In the following we will first discuss some characteristics of texts that can be utilised in computer-aided translation. This discussion is followed by a description of the empirical material underlying this study, namely an English-Swedish translation corpus. We then outline in more detail the various tools and methods that are being studied in the project. These tools and methods fall into four categories: (1) diagnostic tools that characterize texts and text-types in terms of parameters that have a direct bearing on the performance and usability of various computer-supported methods of translation; (2) alignment tools that establish correspondences between source and target texts, on various levels such as chapters, divisions, headings, paragraphs, sentences, phrases and words; (3) data

[1] The project is entitled “Linguistic Engineering for Generation and Translation of Documentation” and is financed by the language technology program jointly managed by The Swedish Council for Research in the Humanities and Social Sciences (HSFR) and The Swedish National Board for Industrial and Technological Development (NUTEK).

acquisition tools that are used to extract data from bilingual corpora; and (4) evaluation tools that are used in the evaluation of translations and checking of properties such as consistency of terminology, variation in phraseology and conformity with a given style-guide. Finally, we conclude with a discussion of what we hope to achieve in the course of the project.

Exploiting structural characteristics of texts for computer-aided translation

A characteristic feature of many types of manuals, in particular computer manuals, is the high frequency of recurrent (or repeated) translation units. This is a fact that can be exploited in translation with quite simple tools, e.g. a computerised phrase-book where recurrent units are stored with their translations. We have found cases where up to 43 per cent of the total text in a handbook was made up of recurring identical sentences and (translatable) phrases (Merkel 1992). The same study showed that 20 per cent of the text in another handbook was made up of sentences that were already translated in the previous version. Recurrent phrases found in both documents comprised 15 per cent of the text. Taken together this meant that 31 per cent of the handbook could have been automatically translated with the aid of a translation memory of sentences and phrases of the previous version of that same handbook. When the recurrent phrases and sentences within the handbook were also taken into consideration, the analysis yielded that 52 per cent of the text was repetitious, either internally or externally.

An example of recurrent sentences and segments in a translation of a computer manual is shown in Fig. 1. Recurrent sentences are underlined and recurrent segments are shown in italics.

Moving Your Computer	***Flytta datorn***
Normal shipping and handling can cause loss of data *from a hard disk*. As a precaution against this happening, *do the following*:	Information *på hårddisken* kan förstöras vid flyttning. Det är därför viktigt att du följer anvisningarna nedan.
Go to your *operating system manual* and follow the instructions to *make a backup copy* of all files and data *on the hard disk*.	Säkerhetskopiera därför innehållet *på hårddisken* med hjälp av anvisningarna i *handboken för operativsystemet*.
Do not use *the backup programs on the Reference Diskette* or *the system programs* to *back up your hard disk. When you have completed making a backup copy*, store it *in a safe place*.	Använd inte *programmet för säkerhetskopiering på startdisketten* eller i grundpartitionen när du ska *säkerhetskopiera hårddisken. När du är klar med säkerhetskopieringen* förvarar du kopian *på ett säkert ställe*.
Remove all media (*diskettes, CDs, optical disks, tapes*) *from the drives*.	Ta ut alla utbytbara media *ur enheterna* (*disketter, CD-ROM, band* etc).
Turn off the computer and all attached devices.	*Stäng av datorn* och alla anslutna enheter.
Unplug all power cords from *electrical outlets*.	*Dra ut alla nätkablar* ur vägguttagen.
Note where all *cables and cords* connect to *the rear of the computer*, *then remove them*.	Anteckna var alla *kablar och nätkablar* är anslutna på *systemenhetens baksida och lossa dem*.
Do not risk injury by moving or lifting the computer by yourself. Ask another person to help you.	På grund av datorns tyngd ska du undvika att lyfta den själv.
Pack the computer and all devices. Cushion them well to avoid any damage.	Emballera datorn. Använd originalförpackningarna om du har dem kvar. Om du använder andra lådor, var då noga med att packa utrustningen så att den inte skadas under transporten.
When reinstalling the computer, be sure to maintain a clearance of at least 51 mm (2 in.) *on all sides* to allow *for air circulation.*	*När du installerar datorn igen* kontrollerar du att det finns minst 5 cm mellanrum *på alla sidor* om enheten *för luftcirkulation*.

Figure 1 Recurrent sentences and segments in a translation of a computer manual.

The techniques described so far are both simple and useful, but we must not neglect the risk that they also bring negative effects with them. A possible danger is that the translation will be too uniform and not varied enough. While consistency is always to be preferred in the case of terminology, there may be situations when different translations should be used, even though the source phrase is a recurrent unit of the source text. In a small pilot study on aligned parallel texts (Merkel 1993) we found a few cases where translators had chosen different translations for the same segment of the source text, but with no apparent harm. In a particular chapter of a computer program manual, 23 source sentence types were repeated between 2 and 19 times. In the target text, 20 of these 23 sentence types had been translated with consistent translations. The 3 sentence types (all of which only occurred twice) which had inconsistent translations could equally well have been translated with a standard translation. An example is given below. In other words, there was nothing special in the context that demanded variation. The most likely explanation to the actual variation is that the translator was unaware of the fact that the exact sentence had occurred at some other text segment.

English: Select the port you want to use.

Swedish 1: Välj den port du vill använda.

Swedish 2: Markera den port du vill använda.

Apart from the dimensions of consistency and variation, one may suspect that the use of fixed translations for recurrent segments, may affect the coherence of the target text. This is a question that we only briefly touch upon in this paper.

Towards more advanced translation tools

A basic hypothesis of our project is thus that an interactive memory-based translation system, i.e. a system providing a terminology database, a database of previously translated units (sentences, phrases and perhaps paragraphs) coupled with a set of tools that derive internal and external recurrence profiles for a given text material would give substantial improvements to the translation process, particularly in speed and certain aspects of quality, such as terminological and stylistic consistency. The aim of the project is to determine the advantages and drawbacks of memory-based systems and to find good designs for them. We also intend to study and develop analysis tools that can predict the behaviour of a memory-based system on a given body of text and other support tools for the translation process. Finally, we hope that the project can suggest ways of integrating memory-based systems and rule-based systems.

An important task of the project is to study the effects on the target text of the translation method used and compare translations made by means of memory-based systems with manual translations. This requires the development of methods and tools for evaluation of translation, which is not an easy task. To test design alternatives we need to consider the format and content of the databases as well as the matching algorithms used in database search. In this connection we investigate different methods and tools for data acquisition and the possibility of making database search sensitive to language-independent information encoded in the source document, e.g. semantic or functional properties of a paragraph encoded as a set of SGML-descriptors[2]. The analysis tools primarily support the identification of translation units of a text body, where strings as

[2] SGML stands for Standard Generalized Mark-up Language. See for example Sperberg-McQueen & Burnard (1994) for details on how this mark-up language is put to use in the Text Encoding Initiative.

well as linguistically more interesting units such as lemmas, terms, phrases, patterns and constructions are considered. The analysis tools can be used diagnostically in several ways. For example, a text profile can be generated showing what parts of it are covered to what extent by recurrent items at various levels of abstraction, and the recurrent items can be checked for counterparts in an existing translation memory. Both kinds of information are relevant for deciding what efforts and resources are needed for the translation of the given text body.

While the purpose of this research is to develop tools for automatic analysis, it should be stressed that the tools usually require human intervention and guidance.

An English–Swedish Translation Corpus

A prerequisite for the exploration and use of bi-texts is that you have one at your disposal. In a Swedish context, English and Swedish are by far the two most common source and target languages to consider. As there were no such English-Swedish translation corpus available at the start of the project, an important part of it is to create one and align the texts at least at the paragraph level.

The text corpus we are collecting consists of five different sets of translations from English into Swedish. There are two different sets of computer program manuals from two companies (1 and 2) where the major difference lies in the method of translation. Text 1 has been translated manually and text 2 has been translated with the aid of a memory-based translation tool. Text 3 will comprise of technical documentation but from another domain than computers and computer programs. Text 4 is a selection of legal documents from the EC which is not technical documentation in the strict sense, but as the production of these translations is done in an industrialised manner, they are interesting as empirical material representative for another text genre. Finally there are two novels in the corpus, which will function as reference material to the rest of the material. The choice of the corpus texts has been made with regard both to the availability of the texts in electronic format and to achieve a balance of different text types and translation methods. Depending on what will actually be marked up and aligned at the end of the project, the size of the corpus will be between four and six million words (i.e., two to three million words per language). In Table 1 the contents of the corpus is described.

	Type	**Size (words)**	**No. of volumes**	**Language**	**Translatio n method**	**Source format**
1a.	Computer manual	500,000	2-3	English	Manual	Word
1b.	Computer manual	500,000	2-3	Swedish	Manual	Word
2a.	Computer manual	500,000	10	English	Translation memory	Book-Master
2b.	Computer manual	500,000	10	Swedish	Translation memory	Book-Master
3a.	Technical manual	100,000 – 500,000	To Be Decided	English		
3b.	Technical manual	100,000 – 500,000	To Be Decided	Swedish		
4a.	Legal documents	500,000 – 1,000,000		English	Manual	Ascii, Word-Perfect,
4b.	Legal documents	500,000 – 1,000,000		Swedish	Manual	Ascii, Word-Perfect,
5a.	Fiction	100,000 – 500,000	2	English	Manual	Ascii
5b.	Fiction	100,000 – 500,000	2	Swedish	Manual	Ascii

Table 1. The Text Corpus

As can be seen from the table the source texts come in a variety of word processing formats: Word, Bookmaster, WordPerfect and pure Ascii text files. This poses a well-known problem to anybody who is trying to mark-up a text corpus consistently. For formatted texts (such as Word and WordPerfect files) it is possible, but not always straightforward, to derive headings, tables and lists, but for pure Ascii files this information has to be handcoded in most cases. Texts written in a more descriptive format, such as Bookmaster, where there are codes for different kinds of elements is relatively easy to convert to your desired format, but the problem lies in what coded information should be kept in the corpus and what could be removed.

We are aiming for a corpus format in SGML along the Guidelines for Electronic Text Encoding and Interchange (TEI P3) which provides a rich set of mark-up possibilities and which is also the only existing standard in corpus processing (Sperberg-McQueen & Burnard, 1994). The process to convert the texts in the corpus to SGML has to be done in several steps:

- Cleaning up original files (for example, replacing graphics with appropriate tags)
- Encapsulating chunks of text into structural units, such as sections, lists, etc.
- Converting paragraph coding into paragraph tags
- Inserting sentence boundaries
- Inserting unique identifiers by means of attributes in each element
- Creating the correspondences between the source and target texts (alignment).

At each step the text has to be validated against the Document Type Definition (DTD) that is set up for the corpus. Apart from headings, paragraphs and sentences the corpus is also marked up for tables and lists. Fig. 2 shows a marked-up sample from the beginning of a chapter of a computer manual.

```
<div1 type=section id=E.OBUP.1>

<head type=h1 id=>OBUP.1.h1>OS/2 2.1 Backup Diskette Package for Preinstalled
Systems</head>

<div2 type=section id=OBUP.1.1>

<p type=text id=OBUP.1.1.p1><s id=OBUP.1.1.p1.s1>This package contains diskettes
and step-by-step procedures that allow you to restore the Operating System/2 (OS/2)
2.1 operating system and preinstalled features to your hard disk.</s><s
id=OBUP.1.1.p1.s2>Use of this package is restricted by specific terms and
conditions.</s></p>

<p type=text id=OBUP.1.1.p2><s id=OBUP.1.1.p2.s1>The enclosed diskettes are for
backup purposes only.</s><s id=OBUP.1.1.p2.s2>Your use of the program materials
contained on these diskettes is restricted to the terms and conditions specified in the
"IBM Program License Agreement".</s><s id=OBUP.1.1.p2.s3>This agreement
entitles you to have only one backup copy of the OS/2 operating system.</s>...</p>

...

</div2> </div1>
```

Figure 2 Sample of mark-up of an English source text in SGML. The elements that are illustrated are for sections (div1 and div2), headings (head), paragraphs (p) and sentences (s). Each element has also a unique identifier (id).

Translation Support Tools

The tools that we consider for translation support can be categorised into four major categories:

- *Diagnostic tools* that characterize texts and text-types in terms of parameters that have a direct bearing on the performance and usability of various computer-supported methods of translation; by applying such tools to a representative sample of texts for a given text-type, a set of text profiles is obtained that reveal characteristics of the text type and that can support decisions as to what kind of computer support should be used in the translation process;

- *Alignment tools* that establish correspondences between source and target texts, on various levels such as chapters, divisions, headings, paragraphs, sentences, phrases and words;

- *Data acquisition tools* that retrieve data from bilingual corpora, which can be exploited in the actual translation process; and

- *Evaluation tools* that are used in the evaluation of translations and checking of properties such as consistency of terminology, variation in phraseology and conformity with a given style-guide.

As stated earlier, we are primarily concerned with memory-based systems, but translations using no computer support or only standard text processing facilities are used as a reference for comparisons.

Diagnostic tools

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

Decision-making (what method or methods are appropriate with a given text-type, and with different parts of a given text?).

System configuration (what data is relevant and how should it be acquired and put to use?).

Post-editing (what effects is the chosen method likely to have on the target text?).

As mentioned in section 2, we have already developed a tool, nicknamed FRASSE (Merkel et. al. 1994), that finds recurrent sentences and phrases in a body of text. Given no other input than a text (and a word list used for filtering purposes), the system retrieves recurrent sentences and phrases of the text and their positions. In addition it provides information on internal and external recurrence rates, that is, how recurrent the text is within itself and compared to other texts.

An example of an output from FRASSE is shown in Table 2. The text which was analysed was a 100,000 word user's guide for a database program. Segments consisting of 3 words or more was searched for without the use of any filters. The result was a list of maximal strings in the source text of which many are useless as translation units.

Table 2 The top 32 segments generated by FRASSE from a computer program User's Guide (without filtering)

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

When we revised this kind of output from the system by hand, we found that a majority of the segments that we removed from the original output actually were segments that contained punctuation marks, ended with phrase-initial function words (such as "and", "the", "to", "in", etc.) or had only one parenthesis character or quotation mark. Therefore we implemented a filter where it is possible to define words that should be stripped at the beginning and at the end of segments as well as requirements on what kinds of characters that should be regarded as pairs (quotation marks, parentheses, etc). Table 3 below shows the result of running the system on the same text as in table 4 with the filtering mechanism on. Of the 32 most frequent segments in the unfiltered result above, 22 have been filtered out, leaving a residue of 10.

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

Table 3 The top 10 segments generated by FRASSE (with filtering)

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

FRASSE can be made more powerful in several ways. A statistical measure which is often used as an indicator of collocations is mutual information (Church and Hanks 1990). Several studies have indicated, though, that high frequency is a stronger indicator than mutual information (e.g. Daille 1994). A better knowledge of linguistic structure may prove more productive. The identification of phrases can be made more precise by shallow parsing of the output on the basis of punctuation marks and function words such as articles, prepositions and conjunctions. Such forms mark boundaries of minimal segments (McDonald 1992) and it can be postulated that the majority of useful phrases have their boundaries coinciding with the boundaries of these minimal segments (a recurrent phrase may span more than one minimal segment, however).

We also wish to abstract from morphological variation in the recurrent phrases. For this purpose it is necessary to combine the existing tool with morphological analysers for Swedish and English.

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

The phrase and sentence lists generated by the current tool can be viewed manually or analysed automatically in order to detect unnecessary stylistic variation, e.g. the prepositions that go with the complements of certain nouns and verbs in examples such as information on vs. information about. A complete description of terminological variation can probably only be obtained on the basis of semantic tagging or an existing terminological data base with a complete coverage of synonyms. In the project we attempt to use the corpus to identify units that have the same translation, concentrating on concepts and constructions where consistency is vital, e.g. in the use of domain terms.

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

Alignment tools

In order to reuse translated material, there must be tools and methods to build up translation memories from existing translations. This is called *aligning* the source and the target texts. Usually alignment is performed on sentence level, but it is also possible to align both larger segments (such as paragraphs) and smaller segments (such as phrases and words).

There are two main approaches to sentence alignment, namely statistics-based alignment and lexicon-based alignment. The statistics-based alignment approach can be based on either character length (Gale & Church 1991) or word length (Brown et. al. 1991). The lexicon-based alignment is used by for example Mariani et. al (1991) to produce correspondences between words by means of a bilingual lexicon. Recently, hybrid approaches to alignment have been proposed, by for example Johansson & Hofland (1994) and Wu (1994), both of which look very promising. Here the idea is that the use of a list of pairs of source and target words with a high likelihood of being translations, complements the statistical alignment algorithm. In Johansson & Hofland's terminology this list is composed of *anchor words* whereas Wu calls the same *lexical cues*, but the basic functionality is the same.

We have modified the alignment algorithm presented in Gale & Church (1991) to produce a simple alignment tool. The system creates translation memories of a source and target text, that is, it links a sentence in the original with a corresponding sentence in the target document. Apart from 1-1 relations, the program also handles 1-2 and 2-1 relations (1 source sentence - 2 target sentences, 2 source sentences - 1 target sentence). This system performs seems to perform well for the texts of our corpus. A first test run on a manually translated text, showed that out of 624 sentences, it failed on only 4 sentences.

The problem with alignment based purely on statistics is that in passages with sentences of roughly the same length, two minor perturbations can cause the alignment of the particular passage to go wrong. In the example run mentioned above, the text comprised of very short paragraphs, with a maximum of six sentences in a paragraph, the risk for misalignment with a purely statistics-based method is small. If the paragraphs are longer the risk of misalignment increases drastically, which is why a hybrid approach is preferable.

In repetitive texts it is possible to use information on recurrent segments for the purpose of alignment. We have outlined such an alignment approach based on length in characters, anchor words and a list of recurrent sentences and phrases which we believe will improve the accuracy of sentence alignment as well as aligning the recurrent phrases (Ahrenberg & Merkel 1994).

Data acquisition tools and system design

Given a translation memory aligned at the sentence level, we now consider how to extract translation correspondences of a more general kind. We first consider the recurrent phrases. When they are proper translation units they ought to correspond to expressions in the target text that are also recurrent. If links between recurrent units can be established, more or less automatically, the identification of other correspondences will become easier. However, it will be necessary to abstract over the strings that FRASSE currently generates. Kaji et al. (1992) describe a system that learns translation templates from bilingual text, i.e. phrasal patterns that contain variables for variable parts, e.g. a pattern of the form “for more information about X[NP], see Y[NP]”. It is to be expected that at this level of description translation correspondences will, as a general rule, no longer be one-to-one. We will study, however, to what extent variation can be predicted when features of the context are taken into account and whether it would be worthwhile to include such features in the rules and in the documents to be translated, e.g. as functionally differentiating paragraph tags.

Moreover, we intend to investigate whether useful correspondences can be extracted also for semantic units, such as word associations and valency frames. This would perhaps make it possible to organise part of the database around concepts, just as a terminological database, although the conceptual units in this case are complexes rather than primitives.

Basili et al. (1992) show how shallow parsing of a tagged corpus can be used to extract simple and complex word associations (e.g. verb – object relations) and Daugaard et al. (1992) similarly extract valency information for Danish verbs by shallow parsing of text corpora. Abstracting further we also wish to consider the extraction of (augmented) context-free syntactic rules from the corpus, combining the information obtained by morphological analysis and phrase identification. These rules can form the basis of a rule-based module (or system).

Evaluation tools

One hypothesis in this project is that the translation method effects the characteristics of the translation. To study the particular effects of different translation methods we will be comparing translations that have been produced by means of memory-based methods to manual translations. This means that we have to use translations that have been post-edited and have different source texts, which makes it hard to reach any definite conclusions. But these studies will serve the purpose of suggesting traits that distinguish the two methods and also suggest improvements of the analysis software.

It could be expected that a memory-based system will support a consistent use of terminology, provided that the database contains the relevant terms. It could also be expected that the translation becomes less varied and that the use of the sentence as the maximal translation unit could make the text less coherent. That is, the context where the example sentence was taken from need not be identical to the context in which it is going to be used, and it is quite possible that the translator does not observe the difference, especially if she is forced by the system to work in a sentence-by-sentence mode.

The study of consistency and variation can take recurrent units as a starting point. Recurrent terms, phrases and constructions are identified in the original and their translations are recorded. Consistency is measured as the number of identical translations. In case of variation, contextual factors that may account for the variation can be identified. By automatically detecting discrepancies in translations of recurrent sentences and segments, the system will highlight the translation variants to the editor/translator. Although objective measures of coherence are hard to come by, we will attempt particularly to study the distribution of indexical expressions such as anaphoric pronouns, definite NPs, indexical adverbs such as however, thus, moreover and indexical adjectives such as same, other, previous. We will also use subjective evaluations of the translations, preferably from professional translators, post-editors and clients.

Summing up

With empirical foundation in existing translations it is our goal to test and develop various strategies and methods that could make the translation of documentation more efficient with the aid of relatively simple tools. Taken together the various tools outlined in this paper, diagnostic, alignment, data acquisition and evaluation tools, will provide the translator with better means of exploiting the information inherent in existing translations and source texts.

The tools will by themselves form a valuable result from the project, but of equal importance is the information that can be extracted from the translation corpus. Questions that we hope to find tentative answers to are: What is the effect of using a memory-based tool for translation as compared with a purely manual translation? To what extent are translations consistent when it comes to terminology and phraseology? What regular correspondences can be identified for a given text type between interesting linguistic units in English and Swedish? And, how large part of a text can be expected to be covered by those correspondences?

References

AHRENBERG, L. & MERKEL, M. 1994. Using Cross-Language Recurrence for Phrase Alignment and Translation Validation. Paper to be presented at the Workshop on Language Engineering on The Information Highway in Santorini, Sept. 26-30, 1994.

Basili, R., Pazienza M. T. & Velardi P. 1992. A Shallow Syntactic Analyzer to Extract Word Associations from Corpora. *Literary and Linguistic Computing, Vol. 7, No. 2*: 113-123.

Brown, P. F:, Lai, J. C. & Mercer, R. L. 1991. Aligning Sentences in Parallell Corpora. In *Proceedings of the 29th Annual Meeting of ACL*, pp. 169-176.

Church, K. W. & Hanks, P. 1990. Word Association Norms, Mutual Information, and Lexicography. In *Computational Linguistics*, Vol. 16, No. 1, pp-22-29.

Daille, B. 1994. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of the workshop The Balancing Act – Combining Symbolic and Statistical Approaches to Language*. New Mexico State University.

Daugaard, J., Kirchmeyer-Andersen S. & Schølser L. 1992. Parsing Large Scale Corpora for Valency Information. Manuscript, University of Odense.

Gale, W. & Church, K. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 177-184.

Isabelle, P., Dymetman, M., Foster, G., Jutrac, J.-M., Machkovitch, E., Perrault, F., Ren, X. & Simard, S. Translation Analysis and Translation Automation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '93)*, pp. 201-217, Kyoto.

Johansson, S. & Hofland, K. 1994. Towards a Parallell English-Norwegian Corpus. (Manuscript).

Karlgren, J. & Cutting, D. 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-94), Vol II: pp. 1071-1075.

Kaji, H., Kida, Y. & Morimoto, Y. 1992. Learning Translation Templates from Bilingual Text. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-92), Vol. II: pp. 672-678.

Kittredge, R., Polguère A. & Goldberg E. 1986. Synthesizing Weather Forecasts from Formatted Data. In *Proceedings of the 11th International Conference on Computational Linguistics* (COLING-86), pp. 563-565.

Marinai, E., Peters, C: & Picchi, E. 1991. Bilingual reference corpora: A system for parallell text retrieval. In *Proceedings of the Seventh Annual Conference of the UW*

Centre for the New OED and Text Research: Using Corpora, pp 63-70. UW Centre for the New OED and Text Research, Waterloo.

McDonald, D. 1992. An Efficient Chart-based Algorithm for Partial-Parsing of Unrestricted Texts. In *Proceedings of the Third Conference on Applied Natural Language Processing,* pp. 193-200, Trento, .

Merkel, M. 1992: Recurrent Patterns in Technical Documentation. Research Report LiTh-IDA-R-92-31. Dept. of Computer and Information Science, Linköping University.

Merkel, M. 1993: When and Why Should Translations Be Reused? In *Proceedings from the Thirteenth VAAKI Symposium on LSP, Theory of Translation and Computers*. Vaasa.

Merkel, M., Nilsson, B. & Ahrenberg, L. 1994. A Phrase-Retrieval System Based on Recurrence. In *Proceedings from the Second Annual Workshop on Very Large Corpora (WVLC2).* Kyoto.

Nagao, M. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn & R. Bernerji (eds.) *Artificial and Human Intelligence,* North-Holland, pp. 173-180.

Sato, S. and Nagao, M. 1990. Toward Memory-Based Translation. In *Proceedings of the 13th International Conference on Computational Linguistics* (COLING-90), Vol. 3: 247-252.

Sperberg-McQueen, C. M. & Burnard, L. 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* The Text Encoding Initiative, Chicago, Oxford

Velardi, P. 1990. Why Human Translators Still Sleep in Peace. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-90), Vol. 2: 383-388.

Wu, D. 1994. Aligning a Parallell English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94),* pp. 80-87.