Towards a Closer Integration of Termbases, Translation Memories, and Parallel Corpora: -A Translation-Oriented View-

Uwe Reinke
Institute for Applied Linguistics, Translation and Interpreting
Saarland University
PO Box 15 11 50
D-66041 Saarbrücken
u.reinke@rz.uni-sb.de

Abstract:

This paper takes a look at how the use of terminological information and bilingual corpora of previously translated texts can improve the performance of translation memories. The focus is on using terminology to support sub-sentential alignment. The author tries to show that the performance of translation memories will not benefit significantly from generalizing the units stored in the memory by replacing terms with variables. Instead, terminological resources should rather be used to support the alignment (and thus the retrieval) of sub-sentential units.
While in many cases current machine-aided term extraction methods will either be too time-consuming or too inexact to produce useful terminology, the results of statistical extraction techniques might be sufficient for generating additional bilingual "anchors" for aligning translation memory units below the sentence level.
The view presented in this paper is "translation-oriented" in so far as the author tries to take into account the fact that any technically possible enhancement will only be accepted by the users if it does not introduce new time-consuming tasks into the translation process.

Keywords: translation memory, terminology extraction, parallel corpora, sub-sentential alignment

Note: This article has first been published in: Sandrini, P. (1999) (ed.): TKE '99-Terminology and Knowledge Engineering-Proceedings. Vienna: TermNet, 527-543.

1 The current state of translation memory technology

In the vast field of multilingual technical documentation, 'integrated translation systems' (ITS)-commonly known as 'translation memory programs' or 'translator

workstations'-are more and more widely used in very different working environments ranging from inhouse language services departments to freelance translators. The two major 'knowledge resources' of such systems comprise a termbase and the reference material (i.e., a translation memory database or a machinereadable collection of previously translated texts). However, until now there has only been a rather low degree of interaction between the individual components, and they take only very little advantage of each other's 'knowledge'. Termbases, for instance, are only used for recognising terms in the text to be translated, while none of the commercial systems employs the terminology available in the termbase in order to enhance the performance of the translation memory component.

A similar situation occurs with respect to previously translated texts that are available in machinereadable form. These are mainly considered as an amalgamation of 'translation units' (TU). No use is made of the terminology that is 'embedded' in them. Although machineaided term extraction is a rather young field of research, some interesting approaches are described in the literature that have not yet found their way into commercial ITSs but have only been implemented in standalone term extraction software and concordance tools.

In the following sections I would like to take a look at how the use of terminological information and parallel corpora [1] can improve the performance of translation memories. I will draw a distinction between explicit and implicit terminological information. Terminological information is explicit if it is stored in a termbase, so that it is directly available to the other components of the ITS. Terminological information is implicit if it is embedded in the parallel corpus but has not yet been incorporated into the termbase. So far, ITSs do not benefit from this latter type of terminology.

The view presented in the following sections will be translationoriented in so far as I will try to take into account the fact that any technically possible enhancement will only be accepted by the users if it does not introduce new time-consuming tasks into the translation process.

2 Making more use of explicit terminological information

Currently, there are basically two suggestions as to how to make more use of the terminology contained in the termbases of ITSs. The first is to use the termbase in order to create more general translation units thus reducing the size of the translation memory. The second is to use terms to improve alignment below the sentence level.

2.1 Using explicit terminological information to create generalized translation units: The "skeleton sentence" approach

In an article published in a special issue of *Machine Translation* that was dedicated to the topic of "new tools for human translators" [Langé et al. (1997) – for references see IAI-2000-Bibliog] propose to use the terminology contained in the termbase of an ITS as a means of creating more general or 'abstract' TUs that replace known terms with variables. The aim of this 'skeleton sentence' approach, as the authors call it, is to improve the recall of translation memories, i.e., their ability to find relevant TUs. The following simple example from the abovementioned paper might illustrate the idea:
(1) Proceed with *installation checking*.
(2) Proceed with *customization*.
(3) Proceed with *X*.
As the terms in italics are stored in the termbase, the translation memory only has to contain the generalized sentence under (3) as a TU. Langé et al. believe that

"a sentence that has been skeletonized to include variable parts is more general, and should therefore be found more frequently in the translation memory than fully instantiated sentences" [Langé et al. (1997:46)].

However, tests of translation memory software have shown that paradigmatic alterations, i.e., changes where both the TU's syntax and its length remain unchanged [2] usually do not cause any retrieval problems (cf. [Reinke (1994)] and [Rösener & Wargenau (1997)]). The only possible advantage of generalizing TUs would then be the reduction in size of the translation memory. Yet, this would only be true for text corpora with a high amount of syntactically identical source language sentences. On the other hand, Langé et al. themselves state a number of problems that are related to the 'skeleton approach', e.g.

- overlapping terms (a sequence of words fits various term entries as in '*Install the receiving antenna support. receiving antenna* vs. *antenna support*')
- variability of terms within a text (identifiying different instances of the same term, e.g., morphosyntactic variants)
- selecting a term in case the termbase contains several synonyms
- agreement problems (adjusting case and gender of nouns and adjectives, subject and verb, etc.)

As Langé et al. mention, these difficulties are typical of any kind of term extraction and term recognition tasks and "they should be addressed at any rate in the terminologyidentification and lookup components of current MAHT products" [Langé et al. (1997:49)].

Yet, in my opinion, the authors do not put enough stress on problems resulting from the differences that might exist between SL and TL representations. The following sentences that are possible German translations of (1) and (2) might illustrate this point.

(1a) *Überprüfung der Installation* fortsetzen. ⇒ *X* fortsetzen.
(1b) Setzen Sie die *Überprüfung der Installation* fort. ⇒ Setzen Sie *X* fort.
(1c) *Überprüfen* Sie als nächstes die *Installation*. ⇒ ??
(1d) *Überprüfung des Leitungssystems fortsetzen.* ⇒ *X* fortsetzen.

(2a) Restliche *benutzerdefinierte Einstellungen festlegen*. ⇒ Restliche *X* (??)
(2b) *Legen* Sie die restlichen *benutzerdefinierten Einstellungen fest*. ⇒ ??
(2c) *Legen* Sie als nächstes die *benutzerdefinierten Einstellungen fest*. ⇒ ??

First, the 'skeleton sentence' approach only works if there is a onetoone correspondence between SL and TL regarding the possibility of replacing terms with variables. Yet, as [Schmitz (1996:200)] points out, "the linguistic representation of a concept can vary between languages from an LSPphrase to a multiword or singleword term" (my translation) [3]. If, for instance, a concept that is represented by a simple noun or by a noun compound is likely to become a complete verb phrase in the TL, the question might arise as to how to generalize the TL part of the TU in case of discontinuous forms (see examples (1c), (2ac)).

Secondly, a solution for treating polysemy and homonymy has to be found. This concerns both 'terms' (cf. examples (1a-c) vs. (1d)) and 'nonterms' (cf. examples (1a,b) vs. (1c) and (2a,b) vs. (2c)).

2.2 Using explicit terminological information for subsentential alignment: The "building block" approach (I)

Langé et al. suggest a second way of exploiting the terminology available in the termbase of an ITS that contributes to the solution of a much more relevant problem in ITS development, i.e., the task of recognising TUs below the sentence level [4]. This task requires a satisfactory solution to two problems: the identification of such subsentential 'units' in the SL and TL texts, i.e., a mechanism for sentence segmentation in SL and TL, and the alignment of the identified units. In this context it should be kept in mind that these fragments could be anything ranging from clauses in compound sentences to more or less complex phrase structures.

Obviously some instrument is needed to assign SL and TL fragments to each other. Current sentence alignment algorithms often rely on the identification of some kind of 'anchor' that connect SL and TL units. These anchors are strings that are identical or very similar on the SL and TL sides of the corpus. Besides formatting tags and punctuation marks, anchors typically comprise figures

(numbers, dates, etc.), proper nouns and socalled 'cognates' (i.e., SL and TL words that "share 'obvious' phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations" [Simard et al.:1992:71)]. It seems to be straightforward to add available bilingual terminology to this list of anchors that can be used during the alignment process. This is actually what Langé et al. suggested as a second way of improving ITSs, but the authors propose using a very shallow sentence segmentation that is only based on a few heuristic rules:

> "we envisage that it [i.e., a phrase TM or 'Building Block TM' as the authors call it; U.R.] will be triggered only in simple cases, for example when the splitting of a sentence into two bricks is made easier by the presence of an unambiguous marker such as a conjunction, or punctuation marks" ([Langé et al. 1997:43]).

The following example derived from a German technical description of a mobile communication system and its English translation might help to illustrate the shortcomings of this approach:

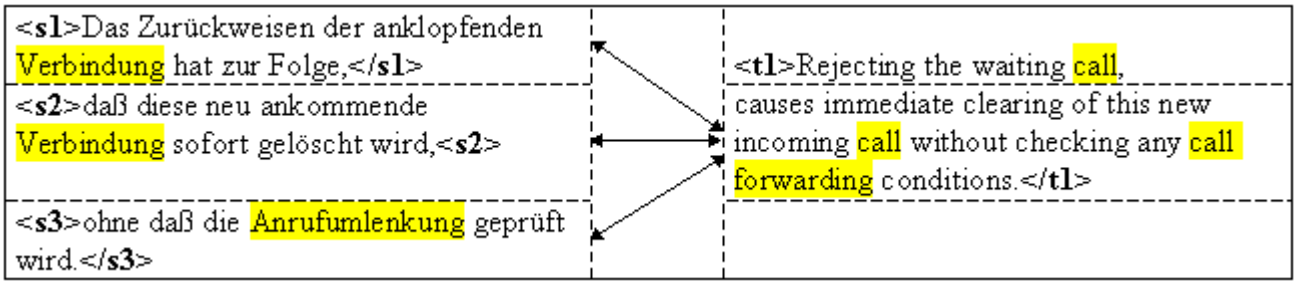
SL: Das Zurückweisen der anklopfenden Verbindung hat zur Folge, daß diese neu ankommende Verbindung sofort gelöscht wird, ohne daß die Anrufumlenkung geprüft wird.

TL: Rejecting the waiting call causes immediate clearing of this new incoming call without checking any call forwarding conditions.

Let us suppose that the following terms are available in the termbase so that they might serve as anchors for aligning SL and TL sentence fragments:

- Anklopfen - call waiting
- Anrufumlenkung - call forwarding
- Verbindung - call;
  connection

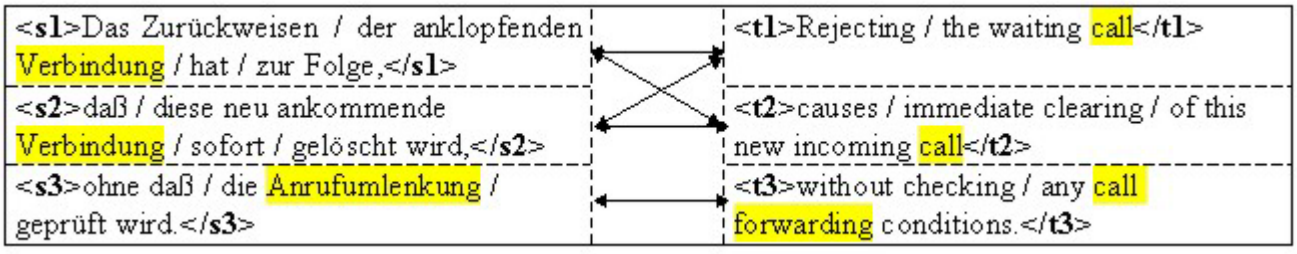
It seems to be obvious that an extremely shallow level of sentence parsing as suggested by [Langé et al. (1997)] would often lead to useless alignments in cases where languages with a comparatively low amount of explicit segment markers are involved. In our example, a segmentation and alignment that follows the suggestions in Langé et al. would probably lead to the following result:



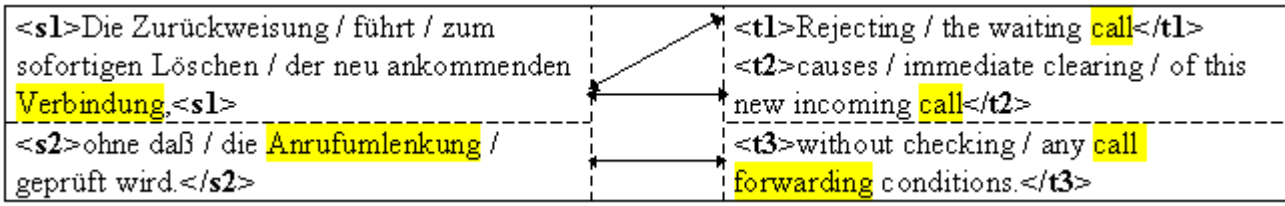
In the following I suggest making use of a conventional rulebased algorithm called PHRASEG that was developed in the 1980s as a segmentation module for

the SUSY MT system [Schmitz (1986)]. PHRASEG can be described as a two-staged parser that builds upon the results of a previous morphosyntactic analysis and partofspeech disambiguation. In the first stage a set of rules is applied to combine word classes that cannot be separated by clause borders. In the second stage these socalled 'phrasings' are then united to segments (clauses). The procedure was implemented for German, English and French and could be easily extended to other languages. For a detailed description see [Schmitz (1986)]. PHRASEG seems to be useful as a segmentation method to support subsentential alignment in ITSs because it offers a level of partial sentence parsing that lies somewhere between shallow and deep forms of sentence analysis [5].

The following table shows how the previous example would be processed when applying the PHRASEG rules. The German part remains unchanged as it contains obvious surface markers. As to the English part, PHRASEG would identify three clauses. Yet, the terms available in the termbase may not suffice for correct alignment.



Moreover, the example also illustrates that there is no real 1:1correspondence between the SL and TL fragments, so that sometimes lowerlevel fragments (separated by '/ ') are related to different higherlevel fragments in the SL and TL sentences (e.g. the verb phrases in <s1> and <t2>). Furthermore, higherlevel fragments need not necessarily align with other higherlevel fragments, i.e. cross-level correspondences must also be taken into consideration. In the following modification of our example a 2:1alignment of <s1> with <t1> and <t2> seems feasible.



3 Making more use of implicit terminological information

There are at least two obvious ways of better exploiting machinereadable corpora of previously translated texts. They can be used either for bilingual term

extraction or for extracting 'anchors' to support subsentential alignment. Evidently, the first option is to the benefit of the termbase, while the second one contributes to improving the performance of translation memories.

3.1 Using parallel corpora to make terminological information explicit: Bilingual term extraction

Feeding a termbase with the terminology required for highquality translation represents a typical bottleneck in the translation process. Only too often tight time schedules and a lack of linguistic data processing skills and/or methods lead to a situation where translators, if at all, only receive very poor glossaries that are not suited to enhance the terminological consistency of the TL text.

Nevertheless, there is a surprising amount of applied research on extracting terminology or compiling dictionaries from machinereadable text corpora that has produced several methods and tools to facilitate this timeconsuming process. Basically, term extraction tools seem to rely on one of the following three general approaches, i.e., a statistical, a linguistic or a hybrid-or "mixed" [Daille (1994)]-approach [6]. With respect to the number of languages involved in the extraction process, we can basically distinguish between monolingual and bilingual methods. The following table is meant as a rough classification of relevant work in this field [7]:

	Monolingual	**Bilingual**
Statistical Approach	[Ahmad & Rogers (1992)], [Ahmad (1994)], [Ahmad & HolmesHiggin (1995)] [Enguehard & Pantera (1994)], [Jacquin & Liscouet (1996)] [...]	[Brown et al. (1993)] [8] [Rapp (1995:97ff.)] [Brown (1997)] [...]
Linguistic Approach	[Ananiadou (1994)] [Bourigault et al. (1996)] [Heid et al. (1996)] [Pearson (1998)] [...]	[Lonsdale (1994)] [9] [...]
Hybrid Approach	[Drouin & Ladouceur (1994)], [Drouin (1997)] [...]	[Boutsis & Piperidis (1996)] [10] [...]
	[van der Eijk (1993)] [Dagan & Church (1994)], [Dagan & Church (1997)] [Daille (1994)], [Daille et al. (1994)], [Gaussier & Langé (19	

	[...]

Although by far not exhaustive, the table might support three assumptions:

1. There seem to be hardly any purely linguistic approaches to bilingual term extraction.
2. There seems to be a tendency towards hybrid approaches that-in some way or other-rely on linguistic information.
3. Most of these hybrid approaches offer solutions for both monolingual and bilingual term extraction.

Bilingual extraction methods are usually based on parallel corpora either computing a probabilistic translation model (purely statistical approach) or calculating associations between potential terms and their translations 'on the TL side'. The latter is achieved by a hybrid approach that usually employs linguistic patterns to identify SL term candidates and either uses the output of a word alignment tool to extract translation candidates (cf. [Dagan & Church (1994)], [Dagan & Church (1997)]) or-more commonly-align the parallel corpus at sentence level and calculate the SLTL associations from the cooccurrences in the aligned sentences (cf. eg. [Daille (1994)] [11].

[L'Homme et al. (1996)] mention three major fields of application for term extraction tools that have their individual user requirements, i.e. translation, terminology, and document management. In this paper we will only consider the area of translation. Yet, even within this field there are quite different purposes for term extraction ranging from automatic generation of bilingual dictionaries for examplebased MT systems to computerassisted compilation of termbases and adhoc projectspecific glossaries for (machineaided) human translation. Thus, even for translation purposes the requirements that a term extraction tool should meet might vary considerably. Compared with the compilation of a termbase or glossary for human usage, a much lower degree of precision might be acceptable when generating a dictionary to support alignment in examplebased MT (cf. [Brown (1997)]).

As to the performance of term extraction tools, at least some of the purely statistical methods seem to produce considerably more noise, i.e., nonterms, than linguistic approaches (cf. eg. [Heid et al. (1996:148)]) [12]. Moreover, tools relying solely on statistics usually cannot extract multiword term candidates (e.g. [Rapp (1995)], [Brown (1997)]) [13]. Linguistic approaches, on the other hand, also have their drawbacks. Apart from the obvious fact that they are language-dependent, they have difficulties in recognizing singleword terms [L'Homme et al. (1996)]. Furthermore, they also seem to produce a releatively high amount of noise [Pearson (1998)]. This might mainly result from an overgeneralization of term formation patterns. Thus, Pearson's work, for instance, provides evidence for the hypothesis that term formation patterns may vary for different subject fields and levels of communication.

All in all, it seems to be obvious that the identification of terms in machine-readable corpora and the compilation of conceptoriented termbases are tasks "that must, in all cases, be carried out by humans during the last stages" [L'Homme et al. (1996:294)]. However, in the translation industry only too often there is no capacity for systematically compiling terminology from available parallel texts. Nevertheless, the terminology embedded in these texts might at least be used to support the retrieval of TUs from a translation memory. Therefore, the following section will look at the possibilities of improving the retrieval of subsentential units by using a simple statistical method described in [Rapp (1995)] to extract further 'anchor points' that support the alignment [14].

3.2 Using implicit terminological information for subsentential alignment: The "building block" approach (II)

Rapp's method of extracting translation pairs requires a parallel corpus aligned on sentence level. For each word of an SL sentence all other SL sentences that contain this word are retrieved. Then, the frequencies of all words in the corresponding TL sentences are calculated. It is assumed that

a) TL words that are translations of SL words appear much more frequently in the extracted set of aligned sentences.
b) Ideally, the frequency of SL words and their corresponding TL translations should be identical.

These assumptions are expressed in the following formula (cf. [Rapp (1995:106)]: absent in original

The likelyhood a_t that a potential word *t* is a translation of a word *s*-or in Rapp's terms the activity associated with *t*-depends on the frequency of its occurrence in the retrieved sentence pairs f_{st} as well as on the relation between its corpus frequency f_t and the corpus frequency f_s of the SL word *s*.

In an experiment, this measure was applied to a very small GermanEnglish parallel corpus of texts that are part of the technical description of a mobile communication system. Each side of the corpus amounts to about 10,000 words. On the one hand this figure might seem extremely small, but on the other hand it might also be a fairly realistic size for many translation projects. In a preparatory stage, the corpus was aligned on sentence level using commercial alignment software, and lemmatization was carried out using the MPRO tool (see above). The frequencies were calculated with the help of WORD BASIC macros.

The following two examples might illustrate that like other statistical procedures this rather simple approach cannot satisfactorily cope with multiword terms. The identification of SL multiword terms is particularly difficult if their indivdual

components appear comparatively frequently as single words or in other word combinations (example 1).

Table 1: Kennungsanforderung ⇔ identity request					
German lemma (*s*) and its corpus frequency (f_s)	**Rank**	**English translation candidate (*t*)**	**Frequency of *t* in the aligned sentences (f_{st})**	**Corpus frequency of *t* (f_t)**	**'Activity' of *t* (a_t)**
Kennungsanforderung 6	1	failure	3	5	2.50
	2	cause	3	8	2.25
	3	correlation	1	6	1.00
	4	previous	2	12	1.00
	5	identity	5	34	0.88
	6	recovery	1	7	0.86
	7	request	5	36	0.83
	8	include	1	8	0.75
	9	interworking	1	4	0.67
	10	identify	1	12	0.50

If, however, the individual components of an SL multiword term only seldomly appear as single word items or as words in other word combinations, then these components rank very high in the list of translation candidates (example 2). In this case the German term together with the first item from the translation table might already be used as an anchor for aligning sentence fragments.

Table 2: Korrelationstabelle ⇔ correlation table					
German lemma (*s*) and its corpus frequency (f_s)	**Rank**	**English translation candidate (*t*)**	**Frequency of *t* in the aligned sentences (f_{st})**	**Corpus frequency of *t* (f_t)**	**'Activity' of *t* (a_t)**
Korrelationstabelle 5	1	correlation	3	6	2.50
	2	table	3	8	1.88
	3	access	2	10	1.00
	4	contact	1	5	1.00

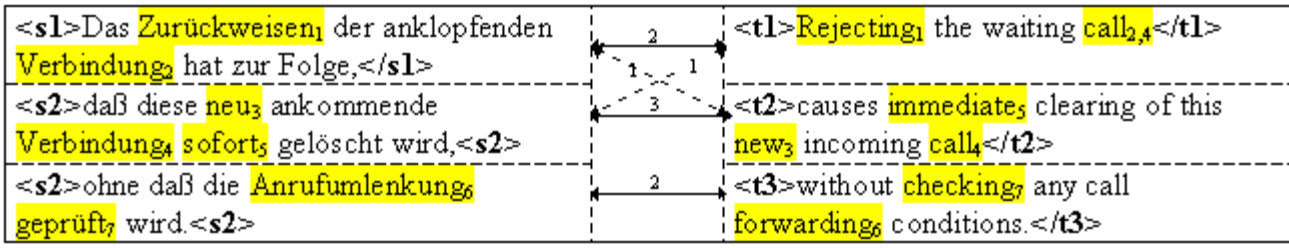
	5	TMSI	8	48	0.83
	6	acknowledgement	1	4	0.80

Taking another look at the former example of aligned sentences and the calculations for translation candidates may underline the assumption that the results of simple statistical calculations could support subsentential alignment.

Table 3: 'Activities' of firstrank TL candidates for all SL lemmata from the former alignment example

German lemma(*s*) and its corpus frequency (f_s)	**English translation candidate (*t*)**	**Frequency of *t* in the aligned sentences (f_{st})**	**Corpus frequency of *t* (f_t)**	**'Activity' of *t* (a_t)**
zurückweisen 9	reject	8	15	4.80
anklopfend 20	accept	9	18	8.10
Verbindung 186	call	225	375	111.60
Folge 4	clearing	2	4	2.00
neu 45	new	40	47	38.30
ankommend 16	incoming	15	21	11.43
sofort 5	immediate	3	8	1.88
löschen 17	cancel	9	13	6.88
Anrufumlenkung 17	forward	15	25	10.20
prüfen 9	check	8	12	6.00

The list of translation candidates shows that some of the calculations produce anchors (e.g., 'zurückwesen' ⇔ 'reject', 'prüfen' ⇔ 'check', and 'sofort' ⇔ 'immediate'), while others are not productive (e.g.,'löschen', 'anklopfend'). In our example the additional anchors improve the alignment of sentence fragments.



4 Conclusions and Outlook

In this article I tried to show that the explicit and implicit terminological knowledge available in integrated translation systems might be exploited to improve the alignment of units below the sentence level and thus enhance the performance of translation memories. While using explicit terminology as anchors in the alignment process seems to be an obvious step, this kind of knowledge often may not be sufficient to achieve acceptable results. Therefore, the integration of term extraction methods into current ITSs and alignment programs might be helpful. Yet, detailed investigations are still needed to find those approaches that lead to reasonable results without being overly timeconsuming.

Another question to be discussed is the way in which subsentential alignment could be integrated into ITSs. Here, at least three different processes must be distinguished:

a) before the actual translation phase:

- the preparation of previously translated material

b) during the translation phase:

- the treatment of the material to be translated
- the treatment of new material that is added to a translation memory.

When using previously translated material to build a translation memory, the alignment tools that perform this task should also contain the algorithms required for subsentential alignment. Depending on text size and processing method, alignment below the sentence level might require a lot of processing time. As long as the output is taken as it is, i.e., without any further manual corrections, this may be acceptable because alignment takes place before the actual translation phase and does not hinder translators from their main jobs.

If a translation memory lookup does not lead to useful results (no match or insufficient similarity between the sentence to be translated and the sentence retrieved from the TM), the user might initiate a search in a database containing aligned SL and TL fragments. Prior to the lookup, morphological and syntactical analysis of the SL material is necessary. This could either be done separately before the actual translation phase, so that the whole SL text is processed in one

step, or the processing could be included in the lookup phase, so that only those sentences of the SL text will be analyzed for which a retrieval of subsentential fragments is required.

When a translator works on a text and adds new data to a translation memory-either by modifiying a suggestion made by the memory or by translating a new sentence from scratch-sentence segmentation and subsentential alignment must be performed before storing the TU in the translation memory. The amount of time needed for these processing steps should be acceptable, as the processing is restricted to a single TU.

Notes

[1] For the purpose of this paper a parallel corpus is defined as a bilingual (or multilingual) collection of texts that consists of at least two subsets, where the texts of subset A constitute the source for translations into one or more other languages (texts of subsets B, C etc.). As to the terminological problems related to the term 'parallel corpus' see [Pearson (1998:47f.)].
[2] In the case of lexical substitutions comparatively small variations in the length of a TU only occur if a term does not consist of the same number of words as the term it replaces.
[3] "Sprachliche Repräsentationen eines Begriffes können von Sprache zu Sprache zwischen Fachwendung, Mehrwortbenennung und einfacher Benennung variieren."
[4] Except for ZERESTRANS (ZERES GmbH Bochum) none of the commercial ITSs offers a possibility for retrieving sub-sentential fragments. However, ZERESTRANS does not apply terminolgical information to the segmentation process but relies on statistical information on part-of-speech co-occurrencies derived from large corpora.
[5] [Schmitz (1986:156)] suggested that together with modules for morphological analysis and part-of-speech disambiguation PHRASEG could perform syntactic analysis in MAHT systems. Unfortunately, PHRASEG is no longer operational. On the other hand, it was one of the few NLP-programs of its time that tried to seperate rules and algorithms as strictly as possible, so that it seems possible and worthwhile to make use of the set of rules as documented in [Schmitz (1986)]. For the morphological analysis I used the MPRO tool developed at the INSTITUT FÜR ANGEWANDTE INFORMATIONSWISSENSCHAFT (IAI) in Saarbrücken (cf. [Maas (1996)]. MPRO builds heavily on the morphological component of the SUSY MT system and includes modules for analysing German, English and French texts. Another formalism designed at IAI for shallow post-morphological processing (cf. [Carl &Schmidt-Wigger (1998)]) was employed for the remaining major part of the analysis. A basic set of disambiguation rules for German and English was provided by IAI while my own efforts were directed towards re-writing the PHRASEG rules. For several reasons, some of the restrictions included in the original rule set could not be implemented. Currently, the rules for German are more or less complete while the English rules still need

further modifications.
[6] This classification is also suggested by [Drouin (1997)] while [L'Homme et al. (1996)] only distinguish between linguistic and statistical appproaches. However, they state "that some systems combine aspects of both strategies" (1996:296) and explicitely refer to some of the above-mentioned research. As there seems to be a tendency towards hybrid strategies, the third category seems to be justified.
[7] For a brief summary of various monolingual and bilingual term extraction methods see [Dagan & Church (1997)]. More detailled descriptions of various statistical measures can be found in [Daille (1994)]. On the major advantages and drawbacks of statistical and linguistic approaches see also [L'Homme et al. (1996)] and [Drouin (1997)].
[8] The authors describe a set of five increasingly complex models for calculating the translation probability of word pairs in bilingual corpora (word alignment). Originally, the models were developed to serve as translation models in a statistical MT system. However, several researchers have applied the approach to term extraction and lexicon generation (cf. [Dagan & Church (1997:103)]).
[9] The author describes a procedure that uses linguistic patterns for separately extracting SL and TL term candidates. In a second step a bilingual concordance tool suggests alignments between these canditates.
[10] Their approach is rather 'on the statistics side', too. According to the authors, the amount of linguistic information is kept to a miniumum and actually only includes lexicon-based lemmatization. Yet, the usage of part-of-speech information is said to be envisaged for future work.
[11] Researchers commonly believe that
"[t]he problems of sentence-alignment, if not entirely resolved, are fairly well understood" [Macklovitch & Hannan 1996:147)].
Tests of alignment tools reveal that these systems still produce some noise when it comes to the recognition and correct treatment of contractions (n:1-correspondences), expansions (1:n-correspondences), omissions (1:0-correspondeces), or insertions (0:1-correspondences) (cf. [Groß (1998)]). Today ITS developers offer tools for sentence level alignment-either as an integral part of the ITS or as a separate piece of software.
[12] Heid compared the results of his linguistic approach to the statistical measure of relative frequency ([Ahmad & Rogers (1992)] and [Ahmad (1994)]) that compares the word frequencies of a corpus of LSP texts with those derived from a 'representative' LGP text corpus.
[13] This is true, for instance, for the methods described by [Brown et al. (1993)] and [Rapp (1995)].
[14] Actually, the dictionary extraction described in [Brown (1997)] served a similar purpose. It cannot be compared to dictionaries that can be found in 'traditional' MT systems. Rather, it was meant to provide an instrument for sub-sentential alignment in an example-based MT environment.