In search of the recurrent units of translation

Lieve Macken
University College Ghent/Ghent University

Translation memory systems aim to reuse previously translated texts. Because the operational unit of the first-generation translation memory systems is the sentence, such systems are only useful for text types in which full-sentence repetition frequently occurs. Second-generation sub-sentential translation memory systems try to remedy this problem by providing additional translation suggestions for sub-sentential chunks. In this paper, we compare the performance of a sentence-based translation memory system (SDL Trados Translator's Workbench) with a sub-sentential translation memory system (Similis) on different text types. We demonstrate that some text types (in this case, journalistic texts) are not suited to be translated by means of a translation memory system. We show that Similis offers useful additional translation suggestions for terminology and frequent multiword expressions.

1. Introduction

Translation memory systems aim to reuse previously translated texts. The basic idea is quite simple. Translation memory systems store source segments together with their translation in a database for reuse. During translation, the new text to be translated is segmented and each segment is compared with the source text segments of the database. When a useful match is found, the retrieved source-target segment pair is provided to the translator. If no useful match is found, the translator translates the segment manually and the newly translated segment is added to the database.

Two processes in the above description are important for fully understanding the potential value and limitations of translation memory systems: segmentation and matching. In translation memory systems of the first generation[1], a segment corresponds to a sentence or a sentence-like unit such as a title, header or list item. The text is segmented on the basis of punctuation and document-formatting information. However, there is a major problem with the idea of using sentences as basic units of translation. Because the matching process is sentence-based, the potential value of the use of a translation memory system depends on the degree of full-sentence repetition of the text to be translated in the database. Consequently, translation memories are mainly used for translating technical documents (e.g. user manuals) or texts with related content (related products) or text revisions.

Several researchers have explored the idea of creating sub-sentential translation memories (Gotti et al., 2005; Planas & Furuse, 2003; Simard & Langlais, 2001). In the domain of machine translation, the current best performing statistical machine translation systems are based on phrase-based models (Koehn, 2009), which in fact assemble translations of different sub-sentential units. The sub-sentential units are sometimes defined as "contiguous sequences of words"; in other cases more linguistically motivated definitions are used.

In this paper, we compare the performance of a sentence-based translation memory system of the first generation with a sub-sentential translation memory system of the second generation. We then compare the translation suggestions made by the two different systems.

The second important process mentioned above is matching. During translation, the translation memory system matches the new source sentence with the source sentences in its database and proposes previously translated sentences to the translator. The system can either return sentence pairs with identical source segments (exact matches) or sentences that are similar but not identical to the sentence to be translated (fuzzy matches).

In traditional translation memory systems, similarity is calculated by comparing surface strings, i.e. sequences of characters. In SDL Trados Translator's Workbench, the similarity threshold ranges from 30% to 99%. The user can change the similarity threshold in order to find the proper balance between precision and recall: If the similarity threshold is too high, potentially useful sentence pairs may be missed (high precision, low recall); if the similarity threshold is too low, the match can be based on high-frequency function words and the proposed translations may be of no use (low precision, high recall).

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

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

For a human it is obvious that the following two sentences are very similar to the example above.

(2) Java® is a registered trademark of Sun Microsystems Inc.
(3)Unix®, X/Open®, OSF/1®, and Motif® are registered trademarks of the Open Group.

However, the translation memory system assigns a fuzzy match of 61% to the second sentence and a fuzzy match of less than 30% to the third. As these examples demonstrate, translation memories contain smaller segments than sentences that can be useful for translators. Bowker and Barlow (2004, p. 4) formulate this as follows: "There is still a level of linguistic repetition

that falls between full sentences and specialized terms - repetition at the level of expression or phrase. This is in fact the level where linguistic repetition will occur most often."

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

We also compare the performance of a sentence-based translation memory system with a sub-sentential translation memory system on different text types. As an example of a first generation system, we use SDL Trados Translator's Workbench[2], which is, according to the LISA Translation Memory Survey (Lommel, 2004), the most widely used TM tool. As an example of a second generation system, we use Similis[3]. According to Lagoudaki (2008), only two commercially sub-sentential translation memory systems are available: Similis and Masterin. Because Masterin only supports English, Swedish and Finish, we opted for Similis as sub-sentential translation memory system.

2. Corpus

Three subcorpora with parallel texts belonging to three domains and three different text types were selected from the Dutch Parallel Corpus (Macken et al., 2007). For each subcorpus, approximately 50,000 words of sentence-aligned parallel text was used to populate the translation memory, and approximately 2,000 words of source-text material was selected as text to be translated:

- The *medical* subcorpus contains European Public Assessments Reports (EPARs) originating from one pharmaceutical company. The texts are rather technical with a clear, repetitive structure. The texts were translated from English into Dutch.
- The *financial* subcorpus consists of a collection of newsletters from a bank that provide financial news for investors. The texts were originally written in Dutch and translated into English.
- The *journalistic* subcorpus contains articles originally published in *The Independent* and translated into Dutch for *De Morgen*.

We expect the highest degree of repetitiveness in the medical subcorpus; the lowest in the journalistic subcorpus.

The manually corrected sentence alignments available in the Dutch Parallel Corpus reveal that a different translation strategy was adopted for the medical and financial documents than for the journalistic texts (see Table 1.). In the medical and financial texts, most of the correspondences at

sentential level are 1:1 alignments (98% and 97%, respectively). In the journalistic texts, the 1:1 alignments only account for 70%; 1:2 and 2:1 alignments for 11%; and null alignments (sentences that were added or deleted) for 16%.

Table 1: Number of different types of sentence alignments as extracted from the DPC

Domain	*0:n*	*n:0*	*1:1*	*1:2*	*2:1*	*n:m*	*Total*
Medical	1	0	1478	12	13	0	1504
Financial	3	7	1425	11	15	2	1463
Journalistic	122	83	881	135	12	19	1252

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

3. Sentence-based translation memory

In our first experiment, we used SDL Trados Translator's Workbench, a sentence-based translation memory system of the first generation. We created three translation memories (one for each subcorpus) and populated the translation memories with the sentence-aligned parallel texts. The obtained translation memories are a reduced version of the parallel corpora, as only unique sentence pairs without empty source or target segment (non-null alignments) are retained. Table 2 presents an overview of the size of the translation memories and the reduction rate.

Table 2: Size of the resulting translation memory actually used by SDL Trados Translator's Workbench

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

A size reduction is seen in all three resulting translation memories, yet only for the medical and the financial translation memories is the reduction due to repetition at the sentence level. In the journalistic texts, the reduction is completely attributable to the removal of null alignments.

We used the analysis function of SDL Trados Translator's Workbench to count the number of exact and fuzzy matches in the respective original source texts. During analysis, SDL Trados Translator's Workbench segments the source documents, compares the segments with the selected translation memory and examines the source document for text-internal repetition. The results are presented in Tables 3, 4 and 5. Different match types are distinguished: text-internal repetitions (*repetitions*); exact matches (*100%*); and fuzzy matches within different threshold intervals (*95-99%*, *85-94%*, *75-85%* and *50-74%*). For each match type, the second column contains the number of segments covered; the third column the total number of words; and the fourth column the percentage of the number of words covered.

Table 3: Analysis statistics (SDL Trados Translator's Workbench) for medical texts

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

Table 4: Analysis statistics (SDL Trados Translator's Workbench) for financial texts

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

Table 5: Analysis statistics (SDL Trados Translator's Workbench) for journalistic texts

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

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

To assess the usefulness of the suggested translations, we pre-translated the source texts with a fuzzy match threshold at 70% and manually inspected the translation suggestions. All suggested translations were considered to be either correct or useful, but the scope was considered limited:

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

From this small-scale experiment, we can conclude that some text types are more suited to be translated by means of a translation memory system than others. A second observation is that the analysis figures should be interpreteted carefully. In the medical texts, the statistics indicate that 30% of the segments recur. However, manual inspection of the sentence-based translation suggestions showed that the impact was considered rather low.

4. Chunk-based translation memory

In our second experiment, we compared the performance of Similis, a commercially available sub-sentential translation memory system of the second generation, on the same test set. Similis is a linguistically enhanced translation memory in that it contains monolingual lexicons and chunkers to

group words into phrases (Planas, 2005). As a consequence, Similis is language-dependent. At present, Similis supports the following seven European languages: English, German, French, Italian, Spanish, Portuguese, and Dutch. Similis can be classified as a sub-sentential translation memory, as it can retrieve matches at the sub-sentential level. Translation memory systems working at the sub-sentential level face more challenges than sentence-based systems. In order to suggest matches at a sub-sentential level, the systems must be able to align source and target chunks (a non- trivial task); and must be able to identify (fuzzy) matches at sub-sentential level and have a mechanism to score multiple sub-sentential matches and select the best match.

In the following section we examine what type of structures Similis considers as chunks and we investigate the ability of Similis to align source and target chunks. In section 4.2, we evaluate the translation suggestions of Similis for our three text types; in section 4.3 we enlarge the size of the translation memories and examine how this affects our findings; in section 4.4 we compare the sub-sentential translation suggestions of Similis with the auto-concordance search of SDL Trados Translator's workbench.

4.1 Quality of sub-sentential alignments in Similis

Similis aligns not only sentences but also chunks below sentence level. In order to evaluate the quality of the aligned source and target chunks in Similis, a reference corpus was created, in which the translational correspondences were manually indicated. For each domain, we selected approximately 5,000 words from the parallel texts used to populate the translation memory.

During the manual annotation task the *minimal* language units in the source texts that correspond to an equivalent in the target texts, and vice versa, were aligned. Different units could be linked (words, word groups, paraphrased sections, punctuation). An example of a manually aligned sentence pair is found in Figure 1. Null links (Ø) are used for source text units that have not been translated or target text units that have been added. More details on the manual annotation process are found in Macken (2007).

It	Het
can	kan
not	niet
have been made	zijn
by	van
a	een
walking	wandelende
dinosaur	dinosaurus
because	aangezien
the	de
scratch marks	schrammen
are	zijn
quite	relatief
delicate	fijn
,	,
with	met
long	lange
grooves	groeven
made	Ø
in	in
the	het
sediment	sediment
indicating	die wijzen op
a	een
large	groot
,	,
swimming	zwemmend
animal	dier
, '' he said	Ø
.	. .

Figure 1: Manually aligned source and target units for one sentence pair

Similis defines a chunk as a syntagma:

> SIMILIS met en correspondance non seulement les phrases mais aussi les chunks (ou syntagmes) avec leur traductions. Un syntagme est une unité structurelle du texte: un groupe nominal ou verbal. Il est défini grâce aux catégories grammaticales des mots qui le composent, et qui sont trouvées par l'analyseur linguistique. Un syntagme est parfois appelé "chunk". (Similis, Guide de l'utlilisateur, version 2, p. 4)

The Edit Alignment function of Similis allowed us to inspect the aligned chunks. As seen in Figure 2, Similis's chunks can consist of sequences of several words, but one-word chunks also occur. Table 6 presents an over-

view of the number of source chunks of different lengths that were aligned by Similis in the three test corpora. The majority of aligned source chunks are relatively short chunks: over 50% consist of maximally two words, and 75% contain maximally three words.

Table 6: Size of the source chunks expressed in number of words and percentage of each type in the test corpus

Size of the source chunk	*Percentage*
1	24
2	32
3	19
4	10
5	8
5-10	6
>10	0

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

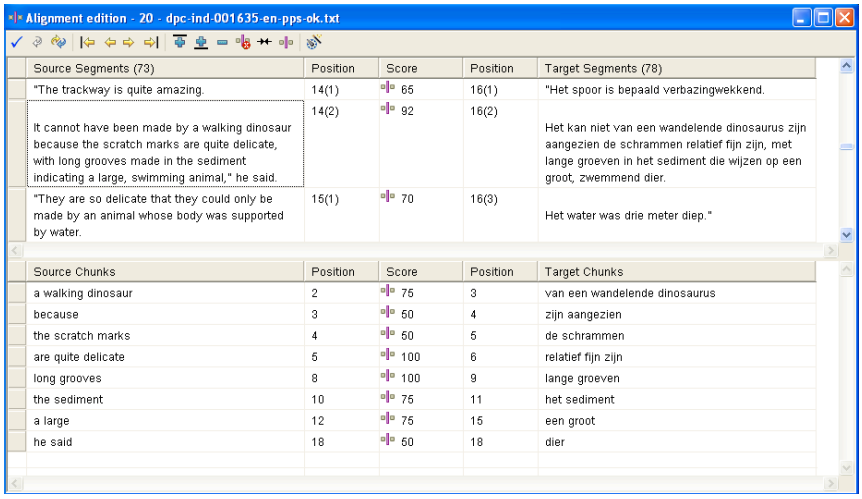

Figure 2: Aligned source and target chunks for one-sentence pairs in Similis

We used the Edit Alignment function of Similis to collect all aligned source and target chunks and compared the aligned chunks with the manual reference.

Each aligned chunk was given one of the following three labels:

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

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

Table 7: Percentages of correct, partially correct or wrongly aligned chunks

Domain	*Correct*	*Partially correct*	*Wrong*
Medical	80%	18%	3%
Financial	70%	25%	5%
Journalistic	67%	26%	7%

4.2 Coverage and quality of Similis's translation suggestions

We used the analysis function of Similis to count the number of exact and fuzzy matches at segment and chunk levels. The results are presented in Table 8. The upper rows present segment matches, which roughly correspond to the statistics given by SDL Trados Translator's Workbench. Minor differences due to application of slightly different segmentation rules and a different calculation of the fuzzy-match scores can be observed.

Table 8: Analysis statistics (Similis) for the three text types: percentage of segments and percentage of words per match type

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

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

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

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

- **Basic vocabulary** if the matched chunk contained only basic vocabulary words.
- **Useful** if the matched chunk and translation suggestion contained some usful suggestion. The match could be a fuzzy match, and the proposed suggestion is not always entirely correct.
- **Wrong** if the proposed translation did not make sense due to alignment errors (see section 4.1).

The results are presented in Table 9.

Table 9: Analysis of the sub-sentential translation suggestions

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

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

However, the percentage of useful sub-sentential suggestions must be interpreted as an upper bound. There are two reasons for this. First, *all* sub-sentential translation suggested were counted, not only the unique ones (e.g. in the financial texts, the word group *de aandelen ~ the shares* occurred several times). Second, whenever the translation suggestion made sense and did not belong to basic vocabulary, the proposed translation was labeled as *useful*. However, the usefulness of most fuzzy matches at sub-sentential level is questionable. For example, for the word group *de Europese nutsbedrijven* [*the European utility companies*], a fuzzy match leads to a translation suggestion of *de Europese beurzen* [*the European stockmarkets*], which is hardly useful, as the translation difficulty is in the noun *nutsbedrijven*, not the adjective *Europese*.

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

4.3 Size of the translation memory

Because it is interesting to examine how the size of the translation memories affects our findings, we extracted additional parallel texts from the Dutch Parallel Corpus. We enlarged the translation memories from 50,000

words to 285,000 words of medical texts, 182,000 words of financial texts, and 289,000 words of journalistic texts. Table 10 presents the analysis results of Similis using the enlarged translation memories. The analysis statistics show that enlarging the translation memory has a positive effect at the level of segment matches for the financial texts: 18.6% exact matches versus 14.5% and 30.3% (all) matches versus 19.3%. Enlarging the translation memory has no effect at the level of segment matches for the jounalistic texts. For the medical texts, there is a slightly negative effect at the level of segment matches, but a positive effect at the level of chunk matches. It seems that if sentences contain fuzzy matches at both segment level and chunk level then the selection mechanism of Similis favours fuzzy matches with the highest threshold regardless of its type. For all text types, enlarging the translation memory has a positive effect on the chunk matches: 23.8% versus 16.3% for the medical texts; 22.2% versus 17.6% for the financial texts and 11.9% versus 6.2% for the journalistic texts.

Table 10: Analysis statistics (Similis) for the three text types using larger translation memories: percentage of segments and percentage of words per match type

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

4.4 Autoconcordance (SDL Trados) versus sub-sentential translation suggestions (Similis)

SDL Trados Translator's Workbench also contains mechanisms to provide the translator with sub-sentential translation suggestions, viz. the auto-concordance search. If no match is found at segment level, the auto-

concordance search retrieves from the translation memory all possible matches on the basis of the segment's lexical items and opens a concordance window showing all matching translation units. Figure 3 presents the auto-concordance result for the sentence "Excessive blood clotting is a problem during severe sepsis, when the blood clots can block the blood supply to important parts of the body such as the kidneys and lungs".

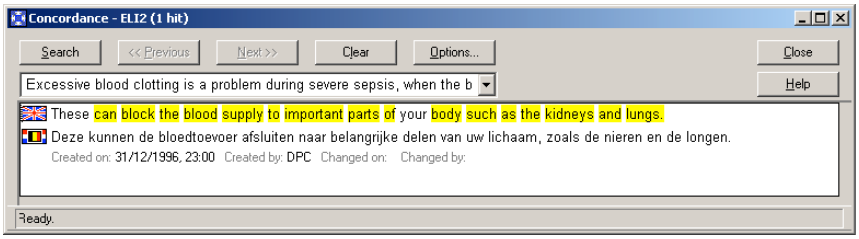

Figure 3: Autoconcordance result in SDL Trados Translator's workbench

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

Figure 4 shows how Similis presents sub-sentential matches to the user. In the sentence to be translated, the sub-sentential matches are indicated by colours. In the example, one exact match (*the kidneys and lungs ~ de nieren en de longen*), and two fuzzy matches (*important parts of the body such as ~ belangrijke delen van uw lichaam* and *severe sepsis ~ ernstige sepsis*) are presented. Contrary to the auto-concordance function of SDL Trados Translator's Workbench, Similis presents the sub-sentential translation suggestions together with the segment matches in the translation environment. Visually, this is less distracting. Moreover, as Similis aligns source and target chunks, translation suggestions below sentence level are presented to the translator and she or he does not need to read an entire series of potentially useful target sentences.

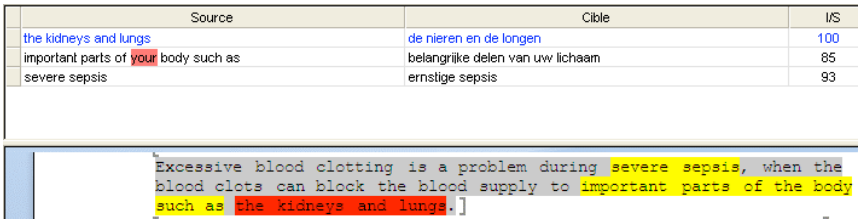

Figure 4: Sub-sentential translation suggestions provided by Similis

5. Bilingual concordance tools

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

Luckily, for such expressions, a bilingual concordance tool such as Paraconc[4], which offers more powerful searches than the concordance function available in SDL Trados Translator's Workbench, may provide assistance. A bilingual concordance tool performs searches on a sentence-aligned parallel corpus. The translator controls the search query and scans the target sentences to locate the translation.

If a bilingual concordance tool is used as a translation aid to solve *lexical* translation problems, relatively large parallel corpora are needed. A large, freely available parallel corpus is Europarl[5], which contains parallel texts in eleven European languages (Koehn, 2005). For the language pairs Dutch-English and Dutch-French, the Dutch Parallel Corpus (Macken et al., 2007) will be available soon.

Figure 5 presents an example of a concordance search for the expression "led the way" in a bilingual corpus. The parallel corpus search offers several Dutch translation suggestions: *de trend zetten*, *als eerste voor iets zorgen*, *het (goede) voorbeeld geven*, *het voortouw nemen*, etc.

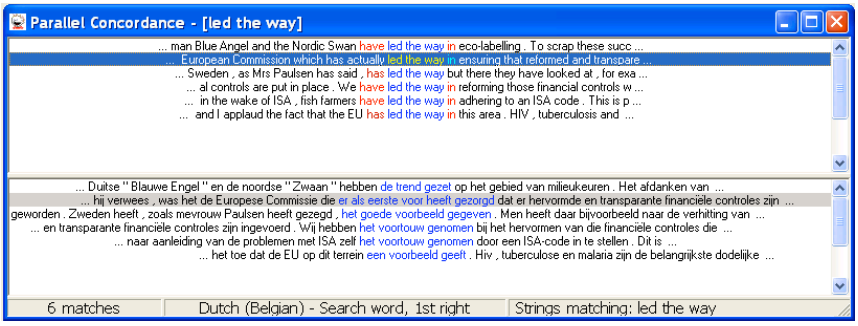

Figure 5: Bilingual concordance window in Paraconc with a contiguous search query

Figure 6 presents a concordance search that was performed for the discontiguous expression "dividend...uitkeren". Paraconc supports wildcards and discontinuity in its search queries, which makes it possible to look for variants of the verb *uitkeren* (*uitgekeerd*, *uitkeert*, etc.) by means of one search query.

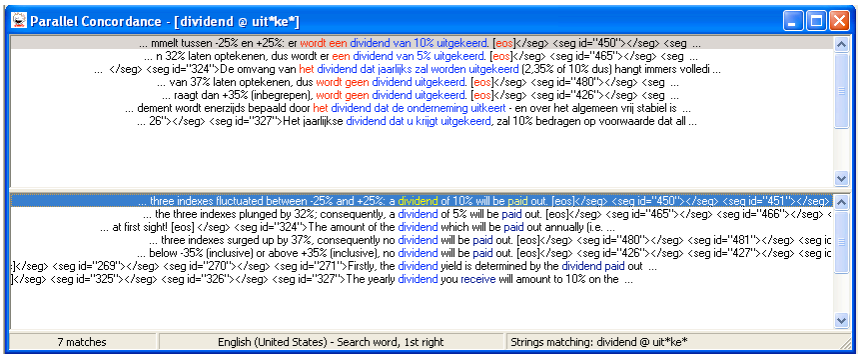

Figure 6: Bilingual concordance window of Paraconc with a discontiguous search query

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

6. Conclusion

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

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

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

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

In order for sub-sentential translation memories to exploit the full potential of translation memories, better word alignment algorithms are necessary so as to improve both precision (the quality of the chunk alignments) and recall (align more flexible units). Ideally, the matching mechanism would also take into account morphological variants, which is a major challenge and a problem unlikely to be solved in the near future.

Bibliography

Bowker, L., & Barlow, M. (2004). Bilingual concordancers and translation memories: A comparative evaluation. In: E. Yuste Rodrigo (Ed.), Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training (pp. 70-83); Geneva, Switzerland, August 28, 2004.

Gotti, F., Langlais, P., Macklovitch, E., Bourigault, D., Robichaud, B., & Coulombe, C. (2005). *3GTM:* A third-generation translation memory. In: Proceedings of the 3rd Computational Linguistics in the North-East (CLiNE) Workshop (pp. 8-15); Gatineau, QC, August 26, 2005.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the Tenth Machine Translation Summit (pp. 79-86); Phuket, Thailand September 12-16, 2005.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Lagoudaki, E. (2008). The value of machine translation for the professional translator. In: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (pp. 262-269); Waikiki, HI, October 21-25, 2008.

Lommel, A. (2004). LISA 2004 translation memory survey: Translation memory and translation memory standards. Romainmotier, Switzerland: LISA. Retrieved June 1, 2005, from http:// www.lisa.org/products/survey/2004/tmsurvey.html

Macken, L. (2007). Analysis of translational correspondence in view of sub-sentential alignment. In: F. Van Eynde, V. Vandeghinste, & I. Schuurman (Eds.), Proceedings of the METIS-II Workshop on New Approaches to Machine Translation (pp. 97-105); Leuven, Belgium, January 1, 2007.

Macken, L., Rura, L., & . Trushkina, J. (2007). Dutch Parallel Corpus: MT corpus and translator's aid. In: B. Maegaard (Ed.), Proceedings of the Machine Translation Summit XI (pp. 313-320); Copenhagen, Denmark, September 10-14 2007. Geneva, Switzerland: European Association for Machine Translation.

Planas, E. (2005). SIMILIS second-generation translation memory software. Proceedings of the 27th International Conference on Translating and the Computer (TC27), London, UK, November 24-25, 2005.

Planas, E., & Furuse, O. (2003). Formalizing translation memory. In M. Carl & A. Way (Eds.), *Recent advances in example-based machine translation* (pp. 157-188). Dordrecht: Kluwer Academic Publishers.

Simard, M. & Langlais, P. (2001). *Sub-sentential exploitation of translation memories*. In: Proceedings of the Machine Translation Summit VIII; Santiago De Compostela, Spain, September 18-22, 2001.

Sparck Jones, K. (1979). Experiments in relevance weighting of search terms. *Information Processing and Management*, *15*, 133-144.

[1] The terms *first generation TM* and *second generation TM* are widely used (Planas, 2005; Lagoudaki, 2008) to refer to sentence-based and sub-sentential translation memory systems, respectively. Only Gotti et al. (2005) make another distinction: first-generation systems are sentence-based translation memory systems without fuzzy matching techniques; second-generation systems are sentence-based systems supporting fuzzy matches; and third-generation systems are sub-sentential translation memory systems.
[2] www.trados.com
[3] www.lingua-et-machina.com
[4] www.athel.com/para.html
[5] http://www.statmt.org/europarl/