[*Translating and the Computer 27*, November 2005 [London: Aslib, 2005]

SIMILIS
Second-generation translation memory software

Dr. Emmanuel PLANAS
Lingua et Machina
6, rue Léonard de Vinci,
BP0119
53001 Laval cedex, France

ep@lingua-et-machina.com
www.lingua-et-machina.com

Abstract

This paper explains how we see the Computer Aided Translation (CAT) new world, introducing Second Generation Translation memories (2gTM).

2gTM use a light linguistic analysis so that the CAT tool is able to deal with the translation of noun phrases and verb phrases rather than complete sentences. This mere feature opens wide the scope of translation memories from 20% up to 80% of the documents that are daily translated in the World.

1. Introduction

In the old times, when the Translation Pangea was wild and hot, in the territory of Tradosaurius, Esdiëlorius and some other Wordfastorius, text documents where considered as a sequence of characters. Ibeëmodocus was already extinct. The horde of translators had to struggle hard for their life while they were to find their path between fuzzy sentences and other 101% matches. A new species, Atrilax (maybe you've seen this before?) introduced the notion of "composed translation", trying to put together patches of previously recorded segments in order to compose a new segment. The attempt was quite "atrevido", but life was still hard.

We propose you a new translation world, where the text is considered as a sequence of words that have a grammatical reality. This new paradigm offers brand new translation automation that we are happy to explain further on.

In chapter 2, we show why First Generation Translation Memories (1gTM) loose an important part of the redundancy of the translated documents. Chapter 3 reminds how men and machine see the document text. Chapter 4 explains shows how Second Generation Translation memories (2gTM) can deal with sub-sentence redundancy. Chapter 5 is a presentation of SIMILIS interface and Chapter 6 goes on with the conclusion and perspectives.

2. Step *back to first generation translation memories*

2.1 A typical translator journey

First Generation Translation Memories (1gTM) collects a series of pairs of sentences (the source and its translation - the target) in a database. It proposes to the translator the pair that is the more similar to the sentence he wants to translate. You all know about the story. Yet, let me take a small example. Here is a sentence to be translated from French to English:

Lingua et Machina présente les mémoires de seconde génération

Now let us suppose that the only sentence pair you have in the memory is:

Les mémoires de seconde génération changent le monde de la traduction.
Second generation memories change the translation world.

A good 1gTM will propose you this only pair with, say a 56% match for you can find in both source sentences "les mémoires de seconde generation".

Remark 1: In a real translation world, the translator would set the fuzzy match threshold to 75% minimum. The only sentence pair of the memory would not even been proposed by the 1gTM tool because it is *below the fuzzy match threshold.*
Now suppose that the translator *really* wants to use *every* proposal from the memory. It has set the threshold to 40% and struggles hard! He presses down the correct short-cut that gives him without waiting more precious seconds the following sentence:

Second generation memories change the translation world.

At this point, please note these two other important points:

Remark 2: In this target sentence, the 1gTM tool does not show the translator the part that is common or not with the sentence that he wants to write (in the target sentence), in order to translate the sentences he has to translate. The translator would just loose time right here identifying the part to be changed.
Remark 3: In order to use this poor fuzzy match, he will have to erase 45% of the sentence. He would loose time again.

2.2 The consequence in terms of text re-use

Here is an extract from a text transcription of a European Parliament session you can find on the web site of the European Parliament. As other sessions, this session has been translated into all European languages. This text is an example of non technical text that is commonly seen as non redundant (0,5% of the sentences are reused).
I have underlined some noun groups and verb groups that could be reused further on the session. I have double underlined the groups that are actually reused in the same text. One can see that even in this same small bunch of text, the reusable groups already represent some interesting percentage of the text.

Duisenberg, President of the European Central Bank. - It is my great pleasure today to present the third Annual Report of the ECB, two and a half years after the launch of the euro and less than half a year before the single currency will become fully visible for our citizens in the form of banknotes and coins.
The year 2000 was a remarkable year for the euro area, as economic growth reached its highest level for over a decade and was accompanied by continued strong job creation. The HICP in the euro area has also, alas, been above 2% since the middle of 2000, mainly owing to oil price increases and the depreciation of the exchange rate of the euro last year. While this increase in the HICP to above the ECB's definition of price stability is not welcome, the ECB cannot and should not avoid short-term price fluctuations caused by such temporary factors. Nonetheless, it is crucial to prevent any spillover of transitory short-term pressures into medium-term inflation expectations. During 2000 the ECB had to be particularly vigilant in

this regard in a context of strong economic growth and given that monetary developments clearly indicated the existence of upward risks to price stability. It was for this reason that we raised interest rates six times in 2000. By doing so, the ECB contributed to ensuring the sustainability of non-inflationary economic growth in the euro area.

On the other hand, let us now consider a technical text that is supposed to be redundant. I have found this text on the help file of my computer operating system :

To connect a printer directly to your computer
Most new printers support Plug and Play, while many older printers do not. The steps involved in installing a printer that is attached to your computer differ depending on whether it supports Plug and Play.

1. Click one of the following links:

- My printer supports Plug and Play.
- My printer does not support Plug and Play.

If you are unsure whether your printer supports Plug and Play, consider the following:

- Does your printer use infrared technology? If it does, your printer supports Plug and Play.
- Consult the owner's manual or packaging of your printer. Most printer manufacturers advertise the fact that their printer supports Plug and Play. Look for Plus and Play on the printer's list of features.
- Check the connector on the end of the printer cable that you plug into your computer.
- If the connector that attaches the printer cable to the computer is a USB connector, then the printer supports Plug and Play.

In terms of sentence fuzzy matches, one can easily see that two sentences are similar:

- My printer supports Plug and Play.
- My printer does not support Plug and Play.

We were expecting this fact since this is a technical text. Nevertheless, there are only two sentences that can be fuzzy reused, while 67 words out of 170 (39%) belong to a group that is redundant in this mere text! Just imagine the kind of redundancy you can get with the entire help file, reusing word groups...

The corollary of this small demonstration is two folded:

- First, 1gTM loose an important part of the text redundancy and consistency when reusing only sentences
- Second, reusing noun and verb phrases expand the type of documents on which Translation Memories can apply.

3. *What is new with second generation translation memories*

3.1. Texts as perceived by humans and machines

In *La structure des langues*[1], Claude Hagège observes that the human perception of a written statement is based on the following three aspects:

- morphosyntactic, which separates classifiable forms (morphology) that can be categorised (nouns, verbs etc.) from their function (syntax). These aspects govern relations between larger sections of the statement (predicate, subject and object).

[1] *La structure des langues,* Claude Hagège, éditions PUF, N°2006, ISBN 2 13 043217 4.

- semantico-referential, which governs the relation between the signified and the signifier (cf. also Saussure[2]) by linking the statement to what it refers to. In particular, this includes the influence of the statement's context (for example, sub-conscious, social or sequence of discourse).
- the enunciative hierarchy that governs the speaker-listener relationship. In Japanese, for example, the speaker will not use the same personal pronoun when speaking to his younger brother *(kimi)* as when speaking to someone who is not a close acquaintance *(anata).*

In an ideal world, we could dream of a computer program able of the same analysis. It is important to realize at first that this is not directly possible in order to loose no time and energy.
As a matter of fact, a program that processes natural language (such as a translation memory tool) will interpret this material on the basis of its own capacities, of which there are seven orders:

- information: texts are considered to be raw information - a series of characters (letters) that can be reduced to binary code at the lowest level.
- morphology: word forms are recognised, for instance in French *mangeaient* = *mang+aient.*
- lexical level: words are identified as lexical units – they are part of a known or inferable lexicon. Each word can therefore be allocated a grammatical category, the ambiguity of which can sometimes only be removed at a higher level.
- syntax: the software understands the syntactical structure of words categorised in the dictionary. Each group is therefore identified, for example, as being nominal or verbal.
- logical-functional: predictive functions, subject and object are analysed.
- semantic: a certain semantic representation of functional groups is deduced with a knowledge that can only be limited.
- context: some data (encoded, for example, in the case of a weather report) allow the program to select one statement rather than another (drizzle, rain, thunderstorm or storm).

I suggest that the human and machine perceptions of text should be seen as coexisting because they share certain common elements.
However, seeking to assimilate each of these seven aspects covered by machines with the three levels of human understanding is an exercise that is a bit too perilous. It is essential to consider that both views are separate.

Once we have clearly this in mind, it is possible to process one or two things. Next section explains how we have succeeded in dealing with these noun and verb phrases.

4. What is new with second generation translation memories

4.1. Why second-generation memory offers translators innovative support

First-generation memory works at the first level: information. It sees text as a code. The word *memory,* for example, is considered different to the word *memories.* First-generation memory is also unable to 'see' syntactical groups in the same way as the translator. The reason that

[2] *Cours de linguistique générale,* Ferdinand de Saussure, éditions Payot, ISBN 2 228 88165 1

SIMILIS second-generation memory offers the possibility of combining two types of programs to work at the morphological-lexical-syntactic level. The software contains a monolingual lexicon for each language processed and algorithms that allow it to analyse and identify grammatical categories of individual words (because the software now works with words rather than a code) and then group these categories of words into word groups. This new capability offers a more efficient translation support tool. To take just one example, SIMILIS can recognise that one group of words (for example, a nominal group) is separate from another group (for example, a verbal group) and then offer the translation for one of the groups as a possible translation later in the text.

5. In practice

5.1. How SIMILIS works

SIMILIS runs a linguistic analysis that sees sentences as a series of syntactical units called 'chunks'[3], which are in turn made up of words. Here is an example of a sentence analysed by SIMILIS: Chunks are indicated into braces.

{Les mémoires de seconde génération} {changent} {le monde de la traduction}	{Second generation memories} {change} {the translation world}
les[le, det] mémoires[mémoire, noun] de[de, prep] seconde [seconde, adj] génération[génération, noun] changent[changer, verb] le[le,det] monde[monde,noun] de[de,prep] la[le,det] traduction[traduction,noun] .[.,mark]	second[second, adj] generation[generation, noun] memories[memory, noun] change[change, verb] the[the, det] translation[translation,noun] world[world,noun] .[.,mark]

The initial analysis looks at the sentence and identifies the chunks, which usually correspond to nominal or verbal groups. The second stage of analysis identifies the basic form of each word of the chunks and their grammatical category.
Since SIMILIS contains lexicons and compares the grammatical structures of source and target sentences, it can find the translation for chunks as long as the languages processed are parallel enough to allow it to do so. So, it holds not only translation units corresponding to sentences in its memory but also those corresponding to chunks. We show in the table below the usable elements for a new translation that come from the analysis of the previous two sentences.

Les memoires de seconde generation changent le monde de la traduction	Second generation memories change the translation world
les mémoires de seconde génération	second generation memories

[3] Cf. Abney

changent	change
le monde de la traduction	the translation world
mémoire de seconde génération	second generation memory
changer	to change
monde de la traduction	translation world

The last stage of parallel analysis run by SIMILIS involves extracting terminology contained in the parallel chunks. This terminology is displayed in the last three lines of the table.

5.2. The benefits of the chunk

Let us come back to the sentence to be translated. Since SIMILIS found the translation for the chunks rather than only for sentences, it can offer a 100% match for the last chunk as opposed to only a partial match for the entire previous sentence.

{second generation memories}

{Lingua et Machina} {présente} {les mémoires de seconde génération}

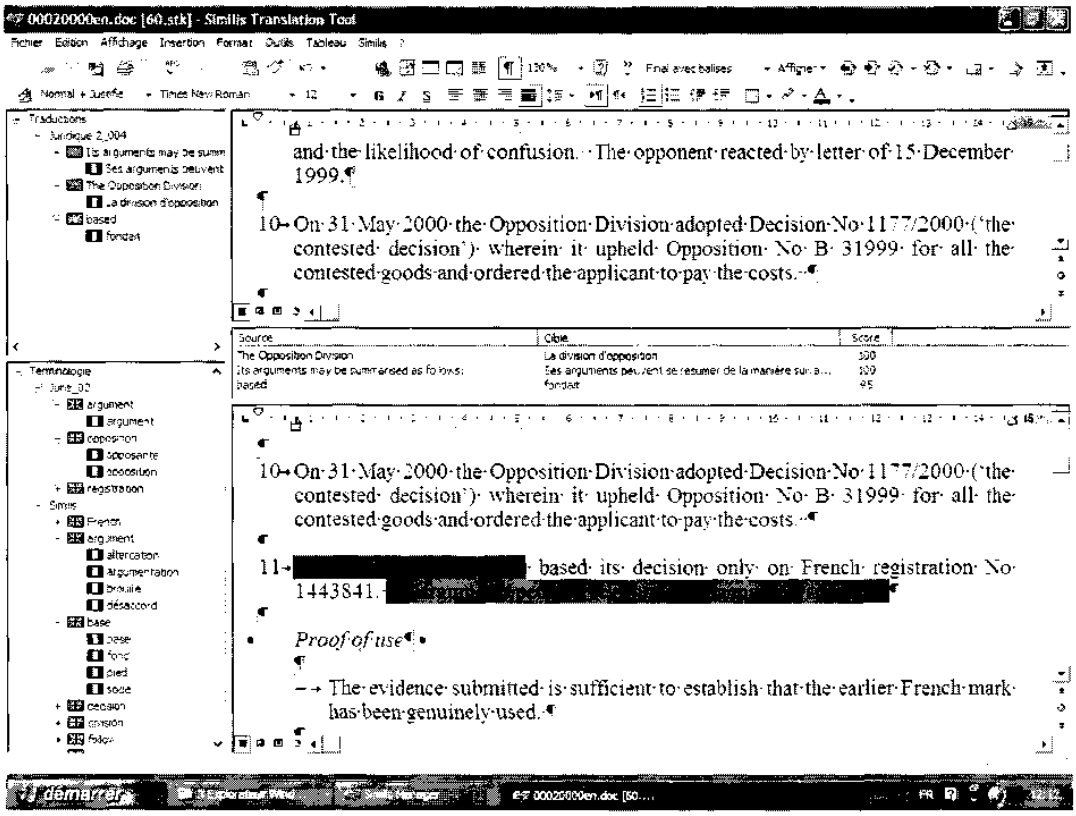
By applying the principle of translation memory to chunks rather than sentences, Lingua et Machina allows translators to use the translation memory concept for a much wider variety of documents.
Moving up from the first to the second generation means that translation memories can now be applied to 80% of documents translated rather than only 20% as was the case for the first generation.

5.3. Importing first-generation memory

When SIMILIS imports a first-generation memory, it not only retains the corresponding segments, but also carries out an analysis of each source and target segment for each translation unit. In this way, SIMILIS builds up the memory by extracting translations of chunks and the associated bilingual terminology.

5.3. Interaction with the translator

This powerful analysis tool is only useful if there is a suitable amount of interaction between the translator and the software. We worked extensively on this aspect with translators so that chunks are suggested in interaction with the translator in a way that is both practical and non-restrictive.
In the interface, the translator translates as though he was not working with a tool. The tool follows the translator, anticipating the chunks he will translate and adapting itself to what the translator produces in order to suggest the chunk or sentence in the memory that best matches his translation.
In the following screen shot, you can see the classical SIMILIS environment. The translator translates in the lowest window where the text will be completely translated at the end of his job. The upper window is a reference window: it shows the original text without any changes. Both windows are synchronised. This allows the translator to refer to it while translating and the proofreader to do his job with both synchronised texts scrolling smoothly.
The central windows shows the segment and chunk proposals coming from the memory. In this screen shot, the translator has clicked on paragraph 11.



At the very moment he clicked in, SIMILIS underlined the whole paragraph in grey so that the translator sees the part of the text he his translating. SIMILIS looked into the translation kit and found that the second sentence has an exact match (in magenta). In the first sentence, he did not find any full segment, but some chunks he shows in red (exact match) and yellow (fuzzy match). The translator can paste these proposals at the right position with a short cut. The left column-sized window shows glossary terms useful for translating paragraph 11. These glossaries have been automatically extracted from previous translations or imported from the translator or his customers own glossaries.

6. Conclusion - what is next

Second-generation memories are about to deeply transform the world of translation by allowing translators to work with a memory of subphrasal units rather than entire sentences. The resulting applications (such as flexible interaction at chunk level and extraction of terminology) will allow our customers - many of whom are already fans - to work more efficiently and quickly.

Now, what will come next? The first next reasonable step would be to introduce a complete syntactic analysis so that it becomes possible to understand anaphoras and solve some difficult sentence segmentation issues.

The second would be to use some semantic analysis so that the machine "knows about" what it is processing. That would allow the CAT tool to propose exact solutions when several are possible. Unfortunately this will only be possible for some restricted domain, and for a restricted set of languages for which a large semantic lexicon (called ontology) will be build-up. People that tell you the opposite would necessary be liars, thieves, or fools...