[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

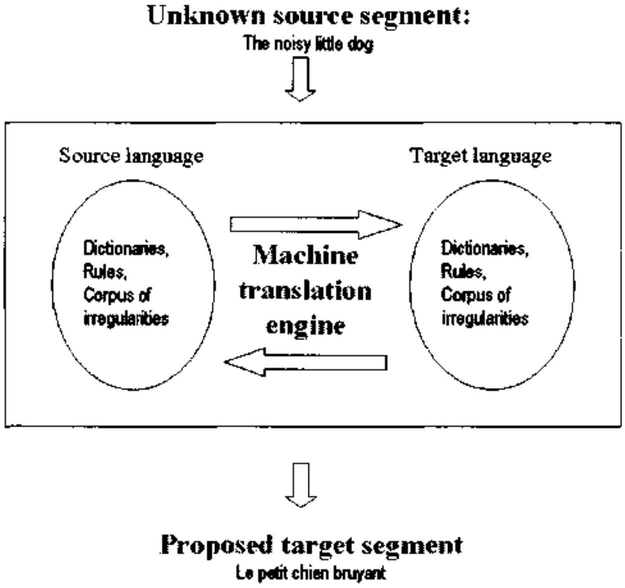
# Automated Translation: The Next Frontier

**Yves Champollion**
www.champollion.net

My subject is a theoretical discussion of the future of automated translation. Can we break through the current Machine Translation (MT) model, and can Translation Memory (TM) help us in doing so?

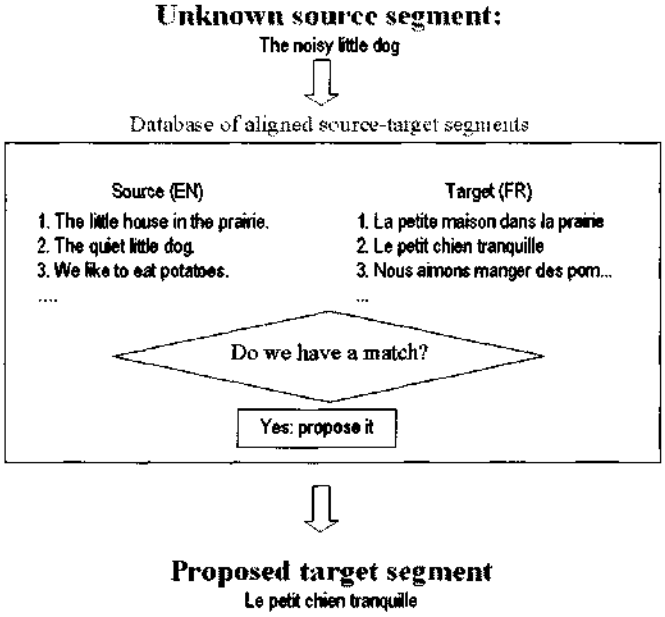
## Introduction

MT applications are pretty simple engines that, knowing the structures of both source and target languages, plus a corpus of irregularities (it would be too simple if languages were perfectly structured), compute a translation as chess software plays chess. Beside the classical Analysis-Transfer-Synthesis model, more recent strategies have been developed, like the statistical approach, aimed at solving ambiguous propositions, but with limited success.

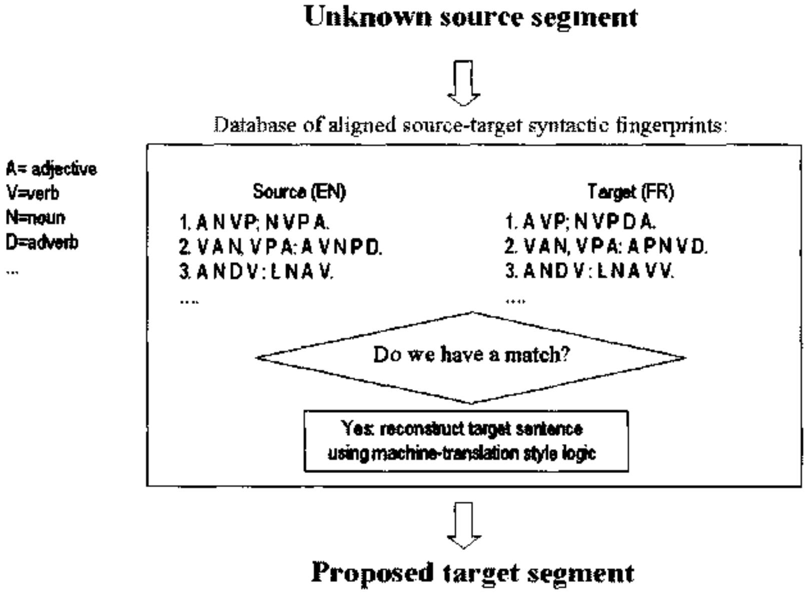


Some further progress is conceivable, by using better dictionaries, more complete exhaustive sets of rules, better corpuses of irregular forms, and better analysis/synthesis software. However, the current MT model has reached maturity and will not get much farther on its own.

Translation Memory (TM), on the other side, is still in infancy. The basic models use databases of existing translations, which we call aligned documents, in both the source and target languages. Reduced fingerprints for each source sentence is stored in an index, and a fuzzy search engine scours the database each time a match is required.


Unknown source segment:
The noisy little dog
Database of aligned source-target segments
Source (EN)
Target (FR)
1. The little house in the prairie.
2. The quiet little dog.
3. We like to eat potatoes.
....
1. La petite maison dans la prairie
2. Le petit chien tranquille
3. Nous aimons manger des pom...
...
Do we have a match?
Yes: propose it
Proposed target segment
Le petit chien tranquille

TM is now beginning to exploit indexes based not on the bare textual contents, but on the structural, or syntactic, contents of segments. To exploit such a database of structural equivalencies, however, TM has to include reconstructivist capacities that are similar to machine-translation algorithms. When TM has not found a "content" match for a given source segment, but has found a "structural" match, the task will be to work on the triangular relationship (source segment to translate, database source segment, database target segment) to propose a target segment, using techniques akin to machine translation, based on dictionaries, rules, corpuses etc.



Obviously, both MT and TM have to join forces in order to bring the next generation of translation automats to life.

## Limits of corpus-based translation

TM is known to perform well in vertical situations, where the new translation that is being undertaken is a sort of repetition of a somewhat similar and previous material. TM makes plenty of sense for corporations that have recurrent translation needs.

Suppose we deal with a project for which there is no previous TM. So, we gather whatever TMs we find and create a general-purpose TM, basically unrelated to the new translation job we have. I call this particular, random TM a "blind" TM.

Blind TMs are known to perform so bad, it's sometimes better not to use them. One temptation would be to make up for the lack of relevance (the lack of "leverage" as we say in the industry) with size - make up for lack of quality with quantity. Could blind databases of huge sizes be of any help? Does size matter? Is TM useful outside the niche of purely "vertical", in-house applications?

In the purely theoretical sense, yes. In the practical sense, no.

Suppose the perfect TM, the ultimate database. It contains all possible and sensical combinations of words a given source language can yield, each combination with a matching translation in the target language. This ideal TM would turn out a 100% match *every* time. This database does not exists yet and I doubt it will ever, but in our thought experiment, let's assume it exists and let's call it UTM for Universal Translation Memory. Each language pair has its theoretical UTM.

What sort of size would a UTM be? To effect a simple comparison, there are far more possible sentences in the simplest of all languages than there are particles in the universe; and if a machine were to produce all combinations of words that make up intelligible sentences (if this were at all possible for a machine), using the fastest machines we have right now, the age of the universe would not be enough. A UTM for a particular language is Utopia, but we will use the *concept* of it for our thought experiment.

Now let's take the largest TM that supposedly exists today on Earth. Let's suppose the Canadian government has nicely archived all English-French translations done in the past 200 years: 100 terabytes of aligned sentences. This looks awesome. On the other hand, compared to the UTM for that language pair, it's ridiculously small, perhaps one billionth of it. And I can predict disappointing results if you use it to translate a recent and trendy Cosmopolitan article on Julia Roberts.

Let's take a detour to a branch of maths that's interested in "Big Numbers" which also deals with mind-boggling numbers and relies heavily on thought experiments. The number Pi (like other transcendental numbers) is supposed to have endless decimals, and they are supposed to follow each other in an unpredictable pattern - you'll never know the next decimal until you've calculated it. Now, is there a possibility for your last name (changed into ascii, or numeric, equivalent) to be clearly written somewhere in the decimals of Pi? The answer is yes, since it's an infinite sequence or random numbers - somewhere, your name is clearly written (and, just as many times, your name is also written with letters in the wrong order - unreadable "noise"). The only problem is spending enough time finding it. But it's there.

To take things one step further, we could ask: is the Bible (Revised King James' version) written in clear form somewhere in this infinite sequence of decimals? Mathematicians are forced to answer "yes", otherwise, the word "infinite" loses meaning. Of course, the same text would be written innumerable times in the wrong order too. That makes a flabbergasting number of decimals in which to look for the right "signal", discarding huge numbers of "noise".

In short, all the knowledge of the Universe is already written - even in each one of us - the only problem being, how can we find it. Here we see the luminous intuition of Socrates, but this is not our subject.

The question narrows down from "Does the right information *exist" ?* to "Can we *find* the right information?". For, if the Bible, or anything, is in the decimals of Pi, the hard question is, how do we find its location. Finding it would take, even with the combined efforts of all men, mice and machines, far longer than the Age of the Universe. Bad news.

To make things worse, the largest number of decimals of Pi computed to date with superlative computers is far from sufficient for my complete name ("y-v-e-s- -c-h-a-m-p-o-l-l-i-o-n") to appear in clear. And King James may have to wait a lot longer.

Getting back to TM and reality. Our question was: can blind TM be efficient? As we have just seen above, the temptation of replacing relevance by quantity is a dangerous one. You can keep adding aligned translations to a database, but you’ll never get anywhere near a UTM, and the leverage increase will be very disappointing. Big Mathematics, however remote from reality it may be, teaches us at least this one point.

In the limited scope of a corporation wanting to build up leverage by archiving its translated material, growing a database does make sense. What I mean, however, is that these corporations should refrain from adding exogenous material, because it will give practically no added leverage - it will surely choke the existing database in the long run. I have seen a few people managing such corporate databases give in to the temptation of "more is better" and having to regret it later.

## Complexity and Machine Translation.

Each language has its own structure. And there is no way, from one language, to guess the structure of another language. So we have an element of both complexity and chaos here and we can ask ourselves: can the recent insights developed by the science of Complexity (also known as Chaos theory) help us here?

Note that this discussion focuses on the challenges of automated translation, not directly on linguistics. I am aware that some languages share similar structures, and that, at least within some languages, some rules allow the modelisation of the evolution of some elements of language.

The science of Complexity made one interesting discovery. Take a system with a set of initial conditions, let it evolve at random, and observe. Observations show that, with a particular set of initial conditions, even if the evolution follows random paths, definite patterns appear, and some sort of stability or equilibrium is quickly reached, at least for a while (like an eco-system). This plateau will remain for some time, then perhaps crumble all at once, then again evolve into another plateau, another equilibrium. But the number of such equilibria, as compared to the nearly infinite number of theoretically possible combinations, is *small* and this is the key point.

The one striking point in the observation of chaotic systems in evolution is that, of perhaps a billion different sorts of possible structures they could evolve into, only but a few of them are evolved into. This becomes clear if you repeat the experiment from scratch many times, as this is now possible with computed "life games"; but it also becomes evident for all who studied the evolution of complex phenomena in the natural or human worlds.

Chaos theorists call this phenomenon "strange attractors". In other words, out of zillions of possible "life forms" a chaotic system could evolve into, only a handful of them are actually evolved into, as if the chaotic evolution was "attracted" to some forms of organization rather than others.

A language can be assimilated to a system in evolution. Even if all languages come from a proto-language, at some point, with different human groups going separate ways, and facing different "special conditions", languages evolved into what they are now (and will continue to evolve).

"Strange attractors" are also found in the realm of linguistics. Even facing different conditions, languages just cannot evolve into "any" form - they're bound to stick to a certain limited amount of possible structures. What is amazing is not how different languages are, but really, how similar they are. This is especially true when one thinks how far apart they could have ended, if "attractors" did not keep them together.

This is where it gets interesting. Even if every language has its own structure, the collection of existing structures is not a random collection, a chaos. Theoretically, it is chaos, but in reality, only a very limited subset of all possible linguistic structures have been developed. Furthermore, there are reasons (which escape us at the moment) for which certain structures are not *possible* at all, and other reasons for which certain structures are *likely.* What research needs to do is to transcend the traditional categories of classical linguistics and move into meta-linguistics with the help of recent breakthroughs in both computational methods and the analysis of complex systems.

Equip a computer with the right software. I mean the sort of software that can spot structures, make deductions from sets of resembling and/or repetitive patterns. A wide array of such software exist: they recognize patterns or forms in graphics. They can identify a face in a picture, which otherwise is merely a sequence of zeros and ones, discriminating between a face and a background. They read maps and guide planes or missiles by taking decisions, avoiding obstacles, choosing an optimal route. They can read vast amounts of data to spot patterns. Other programs, known as fractal engines, can take a hard look at a vast collection of elements, then work out a series of equations, or rules, that can render the same collection of elements in the same precise order: they "structured" - in the mathematical sense - what is seemingly random data and can from then on "reconstruct" it. Even the apparently borderline activity of search for extra-terrestrial life has designed surprisingly accurate algorithms that can differentiate, in an apparently totally random set of data picked up from cosmos waves, "signal" from "noise", where "signal" is anything that could be produced by an intelligent life form. No E.T. has cropped up yet, but the algorithmic that differentiates signal from noise has benefitted from it.

This sort of software, correctly setup and unleashed on vast tracts of bilingual material, will bring out patterns that are invisible to the human eye or brain. Now repeat the experiment, not just on a bilingual corpus, but on a multilingual corpus, ten or twenty languages wide, and you're bound in the end to have an emerging set of structures that, even if they're difficult for humans to grasp, will nevertheless represent some hidden, underlying patterns - and patterns there must necessarily be, as we saw in the discussion about chaos and complexity.

## Synthesis

At this point, one may wonder where all these high-flying and abstract consideration do lead us, practically speaking.

As explained in the introduction, current models (TM and MT) must unite. I do believe that machine translation must keep its central position as the rule-based computational foundation. But recent trends indicate that there is an increasing tendency to complement rule-based (or algorithmic) translation with increasingly vast corpuses: corpuses of expressions, of idiomatic constructions, specialized glossaries that are automatically loaded when a certain context is detected; and most of all, translation memories, whether they be content TM or structural TM.

Corpuses there will be, in vast numbers - they keep coming in droves. We could easily drown in such an ocean of data. For the programmer, the daunting task is to actually leverage results out of this huge mass of data.

When faced with an awesome quantity of data, one answer is to actually limit this mass of data by compressing out of it what is found to be redundant. Lots of data may not look redundant at first look - but the more intelligence is put to work, the more redundancy is found, and the better the database can be streamlined. There is a direct relationship between the amount of intelligence and the bulk of data: intelligence shrinks data.

Take a "structural" or "syntactic" translation memory extracted out of a huge database of sentences: chances are, 90% of sentences are described by a minority of structures, the rest being incompressible: rare structures of speech, unfinished or broken sentences etc, which we could call "noise" for practical purposes. We have a so-called power law distribution: a fairly small collection of syntactic structures actually describe a majority of sentences.

After discarding "noise", you have shrunk your "content" database by making a structural replica of it - a structural database, perhaps one-hundredth the size of the "contents" database. And if you later fall on other data mines of aligned corpuses, you will find that (after eliminating the minority of "noisy" sentences), they will practically not increase the size of the "structural" database.

Note that we don't discard the "content" database (mass storage does not cost much these days). Our structural database, however, now works as an index pointing toward it. In front of every entry in the structural database, we keep an index that points to the various implementations (original forms) of that structure, in the "content" database. But the key point is, when searching for matches, we first search for a *structural* match, i.e. on a database that is very limited in size; and with an intelligent indexing system, we immediately find, then retrieve, a top-ten list of "content" entries that have the most similarities with our original sentence.

As said above, we have left out "bizarre" sentences (sentences that are deviant, or do not fit in the essential core of structures for a given language. They are unfinished or broken sentences, obscure techno-jargon, oddities etc). Our aim, in any case, is to raise the overall productivity of computed translation, not to produce the perfect translation machine (which I believe, is nowhere near us).

From this point on, a rule-based, or algorithmic machine-translation program can produce translations in a much more accurate way by complementing the classical Analysis-Transfer-Synthesis method with a triangulation method using the source segment, and sets of paired approaching matches of the same structure. This structured-corpus approach is valid for sentences that are rather long, but this is precisely where classical machine translation fails to deliver.

This method is but one way to complement machine translation with very large corpuses, all the while maintaining decent speed. It foresees the use a structural, or syntactic, compression of existing translation memories. But one can perhaps, as was pointed above, imagine a totally different "syntax", a set of rules (which perhaps the human mind cannot comprehend) that have been extracted from large databases of aligned material by computer automats, as explained in the section "Complexity and Machine Translation". These rules may look at first totally wild and bizarre, but experience will surely show that they can yield results - after all, these rules are computed out of sets of translations made by humans, and are precisely translation-oriented. Shrinking a large database with such rules and using it will certainly lead to interesting results - but this is a vast enterprise that would require serious research.

i

[i] Bibliography

"Revisiting the Edge of Chaos: Evolving Cellular Automata to Perform Computations" by Norman H. Packard, Santa Fe Institute Working Paper 93-03-014, 1993.

"Life at the Edge of Chaos," by Christopher Langton, in Artificial Life II: Proceedings of the Workshop on Artificial Life (New York: Addison-Wesley, 1992), pp. 41-91.

" The Origins of Order: Self-Organization and Selection in Evolution", by Stuart Kauffman (New York: Oxford University Press, 1993).