Technical Seminar on Copyright, Intellectual Property and Translation Tools
Barcelone, 13 October 2007

The free, universal TM: are idealism and pragmatism compatible?

Yves Champollion, CEO, Wordfast LLC. www.champollion.net

Abstract:

Building a set of public, free, very large translation memories supporting all language combinations is a challenging project. The VLTM project aims at offering translators a repository of TM in all languages, made available for free, with a search-engine approach. The author will explore the technical, entrepreneurial, and deontological aspects of this idealistic project, discuss its feasibility, and present the current state reached by the project.

Translation memories are now widely used in the translation industry at all levels: clients, middlemen (agencies), and translators. Like all technology breakthroughs, TMs are initially met with a mix of praise and suspicion. Translators, for instance, are famous for entertaining a love-hate relationships with this new technology.

This relationship echoes the relationship we had with the first word processors. Word processors were viewed as a great leap forward when compared to typewriters. But then, clients expected a translator to produce more output and less typos with a word processor than with a typewriter. So the translator was feeling, I spend money buying a word processor, training myself to type, and the result is - clients expect me to translate not one thousand, but two thousand words a day... for the same money, and they go mad at the first typo.

CAT tools bring pretty much the same pattern.

The business of globalized TMs

This attraction and repulsion feeling is easily understood. Translators invest money and training in translation tools, having fallen prey to the famous advertising line "*Never translate twice the same sentence*". After paying an arm for the tool, and a leg for training, they discover that in the end, the big guys (the agency or the client) are the ones who own the TM and that the advertizing slogan "*Never translate twice the same sentence*" can also be taken to mean "*Never be paid twice for the same sentence*".

This perception is of course not entirely correct - I am cartooning the situation to make a point. Translators are knowledge workers who specialize in translation. In that form of industry, TM is a form of capital, and as such, it naturally tends to **concentrate**. This is a normal tendency in any market, whatever its political flavour. Campaigning against that would be like campaigning against the weather. It is going to happen, the question is - are we ready?

The concentration of immaterial forms of capital nevers lasts very long, however, especially in an age when technology renders information ubiquitous, and free-flowing. Web servers offering free access to large quantities of TM are bound to appear.

The VLTM project (Very Large Translation Memory - open, free, instant, very large, universal TM for all) means that this is happening right now.

Figures

Until now, most translation memories of large sizes were the property of translation agencies, or of large corporations equipped with their own in-house translation department. The individual translator is ill-equipped to maintain large repositories of translation memory, and he/she is not likely to have any large TM in the first place. An average translator working full-time will produce an estimated 50,000 translation units per year.

Large corporate TMs begin at roughly 2,000,000 (two million) translation units per language pair - 40 years of work for an individual translator. A worthwhile large-scale, all-purpose TM server should offer 20 to 50 times that much to cover a wide spectrum of subjects; and this figure should be multiplied by the number of supported languages.

Fre-lance translator	50,000 TUs
Corporate TM	2,000,000 TUs
Public TM	100,000,000 TUs

One may say that the technological challenges are enormous. But one should also note that general-purpose search engines have already tackled the problem and have working solutions, where huge quantities of information are made available for free, with quasi instant response times. The question really is not *whether* such public, free repositories of translation memories are feasible, but *when* they will appear.

The technology of globalized TMs

There are basically two methods of making TM content available to a wide public of potential users. One is the offline method, the other one is the online method. The offline method is a "Pull" approach, the online method is a "Push" approach. There are current attempts at implementing the offline method. A TM broker will market, license, or otherwise distribute the TM you need. Some websites will let you look up one sentence at a time and find a matching sentence in the required target language, or conduct a concordance or context search.

The online, "Push" method means the translator's desktop tool is directly connected to a TM server over the web. The translator uses is/her usual translation tool to connect to the remote TM, with the option of using a local TM as well. This scheme is used by corporations to have translators teamwork and share a common TM over the web. When the server has a translation to propose, the proposed target segment is automatically brought to the translator's attention, without the translator having to go and get it. This is the "Push" technology.

What keeps these two approaches marginal is that they are both very expensive. A corporate translation memory server based on the "Push" approach costs a little fortune. Buying brokered TMs is also an expensive solution, and it is, in effect, rarely used. Let us compare the situation to the more general question of data gathering before Google appeared (does anyone remembers?). Up until the mid-nineties, gathering information meant having an employee or a subcontractor work for days to gather information. The advent of free-for-all search engines that delivered instant results on huge masses of data sounded incredible at first. Then it sounded like an interesting idea set up by young talented visionaries. Then it turned into technology that worked. Then it proved a viable business model. Now it takes the whole industry by storm.

You don't need to be a prophet to predict that the availability of TM will follow a similar path. Retrograde minds will argue that having huge repositories of TM made available for free is not feasible because of Intellectual Property issues, price, speed issues, corporate policies, whatever.

The VLTM project is now 9 months in existence, and it provides translators all over the world an instant, free, anonymous access to a repository of TMs in any language pair (with 16 languages already having a few million translation units each, and other languages growing quickly).

Here are the main issues involved in building such a project, and how they are addressed:

Gratuity

The VLTM project, like major search engines, does not need any prior registration, is anonymous, and of course, totally free. No advertizing is tied in the service right now, although that could be an option to keep the project running for free in the long term.

Size

The public part of the VLTM does not record the translations done by translators to respect the confidentiality of their work, so one may ask - how does the VLTM grow? Translators, agencies, corporations, NGOs, and institutions can donate TM content that is not deemed confidential. Donated TMs are screened, then added to the relevant language pairs in the VLTM. Another, more radical growth method is outlined in the following section on "Intellectual Property".

Intellectual Property

As mentioned in the previous point, the VLTM grows from TM donated by individuals or corporations who share the vision. Donated TM is considered public domain.

One other source in the future can be the addition of TM produced by aligning multilingual material gleaned over the internet. This may have some people raise an eyebrow, so we'll devote some time in discussing the issue.

Google (and other search engines) used to only direct users to sources of data, most of which constitute Intellectual Property. These search engines acted like directories, not stockists, of other people's intellectual properties. They offered pointers on information sources, like the Yellow Pages. Then Google and the like started to cache content, which means they actually started to copy and store content, most of which is copyrighted, to then deliver it on demand. This has sparked some controversy, which most of you are aware of. As a result it has become commonly accepted that publishing material on the web is tantamount to making it available to the public for free. It does not mean relinquishing a copyright, but it still means making it available to a general, multinational public for free.

Distributing aligned material gleaned on the internet is far less a copyright issue than Google storing and distributing content. The reason is simple. A TM engine will not deliver an entire work, like a book, not even an article, not even a whole page. A TM engine serves, at the most, one isolated sentence at a time, and a couple neighbouring sentences for context. A TM engine only **quotes** what another translator has done in a similar situation. And this does fall under the accepted *right of quotation*. Whereas Google can serve you an entire document - and apparently nobody minds - a TM engine will only quote something like: "You're about to translate sentence A, and for your information, by browsing the public web, I have found that a similar sentence B was translated into C".

In fact, prime translation memory content already exists and is already distributed freely. Use Google to find the same version of a User Manual in two languages, and here you are. The VLTM only makes the process easier and faster; and you won't retrieve the entire document - only sporadic quotations from it.

The VLTM project has not started aligning the web, but it is seriously considering doing so. For two reasons. One is to offer vast resources to translators. The other one is to re-empower translators. Commercial corporations or private agencies are building translation memories for their own commercial or private purposes - and there is no problem doing so. This paper is not an anti-business manifesto: business is all right, and the concentration of capital is a natural tendency. But translators need to have translation content made available to them in a direct, non-commercial, unhindered, free-flowing way.

We should not forget that there is a great need for pro-bono translation, for NGOs, for minority languages, for pure cultural or academic concerns, where the rules of big business do not apply. Translation memory is a form of capital. Even if some concentration of capital in the corporate world is legitimate - we do live in a liberal world - we should still make sure that some of it, if not most of it, is public, shared, and free.

What would our world be if wealthy philanthropists had not built and funded thousands of public libraries centuries ago?

Quality

One legitimate concern is the quality of the translation units stored in the VLTM.

A liminary remark is that Translation Memory technology, like all computerized processes offering assistance to human translators, does not translate. *Translation Memory does not translate* - this is worth repeating. Translation memory offers assistance to the human translation process, and it does so by **quoting** (emphasis added) what another translator has done in a **similar** (emphasis added) situation. When a translation tool brings up a match, the TM engine actually means "I have a record of a colleague of yours faced with a similar sentence, and this is what he/she did". Period. The translator must evaluate how much of that proposition can be re-used: *all* of it, *none* of it, *some* of it. Most translation tools offer the possibility to check what the original context was, some offer to penalize 100% matches that are drawn from a different client, from a different job, etc. Now, even so-called "golden" or "in-context" 100% matches can be wrong. In the end it all boils down to the wisdom of the translator. Although this point relates to TM in general, not specifically to the VLTM, it is a point worth making, because the VLTM project is about making TM more popular and widespread. Although TM quality is an important factor, quality is not a make-or-break element. Most corporate so-called "gold" TMs do contain typos, spelling errors and outright translation inaccuracies. But *most* of their content is valuable, and that is why they're considered precious.

More important: the kidney effect. Although the VLTM does not record what the translator locally translates - to protect confidentiality - it nevertheless records one important feedback. If the VLTM serves an exact match, and that match is re-used ***as is*** by the translator, the VLTM increases that particular TU's "reuse counter". This is like a translator voting for a particular translation - saying he/she agrees with that translation unit. Conversely, we can have a "no-reuse" counter for TUs that are repeatedly offered and repeatedly turned down (rejected or edited) by translators. This means we have content-rating at the VLTM, like page-ranking in other search engines, which makes sure that content value is increasing over time.

This scheme is of course not a panacea that will miraculously make VLTM content perfect in the end. It only shows that the VLTM is not a flat, static repository of translation units, it is a dynamic, bettering system that is constantly tied to human activity, and constantly re-evaluated. The VLTM only aims at being an *aid* to the translation process, not replacing translation.

Ease of use

The VLTM project is a typical "Push" technology. What made the Blackberry a huge success in the US is that, unlike having portable mail-enabled systems where you have to make a connection, log in to your ISP, click "Download mails", the Blackberry has all your mails right there anytime you open it. The VLTM is not a web site you go to and look up content, it pushes content to the translators' tool without him/her having to go and get it. This small technical difference makes a huge usability difference and in the end - it's the difference between sailing and capsizing.

Furthermore, use of the VLTM does not change the work habits and usual workflow of translators - it seamlessly integrates in it. There is no learning curve, no retraining involved. It is, as the saying goes, an unobtrusive plug-and-play technology.

Confidentiality

The VLTM has two parts: public and private.

The private part of the VLTM is a large, public-domain translation memory that is used by everyone on a read-only basis. Nothing of the local translators' work is being fed back into the VLTM. The only thing a translator risks by using the VLTM is receiving free "matches". Confidentiality for translators is utterly preserved. The public part of the VLTM is read-only.

Translators can also create password-protected workgoups that allow them to use a sub-part of the VLTM in read-write mode. This simply means that translation units written into the VLTM by a particular workgroup of translators are shared ***only*** among members of that workgroup. Passwords are long and complex enough to make sure they cannot be cracked, and client-server communications are encrypted and multi-packeted.

Informal groups of translators, but also agencies or corporations can use a private workgroup within the VLTM, enjoying the benefits of a web-shared TM without the huge costs generally associated with TM servers. This perhaps needs repeating: corporations and agencies can now set up private over-the-web translation memories that enable global workgrouping among translators without the cost of ownership of a TM server, without the need to host an actual server, without the need to retrain translators, without the need of a localization engineer, etc. And while translators are sharing a private, specialized TM over the web, they still have the benefit of the large public VLTM and its free content.

Power and scalability

The VLTM can be infinitely scaled up. All it takes is adding actual servers to increase the VLTM's capacity - no corporation or institution will ever outgrow the VLTM. We built search-engine strength into the VLTM project right from the start.

We have started programming an automat that works like most search engine crawlers that harvest web pages, but whose purpose is to find alignable material (multilingual web sites or material), and effectively align the harvested material to feed the VLTM database.

Versatility

Translators can set up their tool to use either the private, or the public, VLTM, or both at the same time. They can also keep using a local TM on their private hard disk as a third source of translation units. They can set up their translation tool to consider matches from the local TM, or from the VLTM, as more valuable by applying simple rules of preference.

Universality

All languages are supported.