Translation Wikified: How will Massive Online Collaboration Impact the World of Translation?

Désilets, Alain

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Translation Wikified: How will Massive Online Collaboration Impact the World of Translation?**

Désilets, A.
November 2007

* published in the Proceedings of Translating and the Computer (29). November 29-30, 2007. London, United Kingdom. NRC 50331.

Canada

Translation Wikified: How will Massive Online Collaboration Impact the World of Translation?

Alain Désilets, National Research Council of Canada
alain.desilets@nrc-cnrc.gc.ca

Opening keynote for Translating and the Computer 29, Nov 29-30, 2007, London, UK.

Abstract

Massively collaborative sites like Wikipedia, YouTube and SecondLife are revolutionizing the way in which content is produced and consumed worldwide. These fundamentally collaborative technologies will have a profound impact on the way in which content is not only produced, but also translated. In this paper, we raise a number of questions that naturally arise in this new frontier of translation. Firstly, we look at what processes and tools might be needed to translate content that is constantly being edited collaboratively by a large, loosely coordinated community of authors. Secondly, we look at how translators might benefit from open, wiki-like translation resources. Thirdly, we look at whether collaborative semantic tagging could help improve Machine Translation by allowing large numbers of people to teach machines facts about the world. These three questions illustrate the various ways in which massive online collaboration might change the rules of the game for translation, by sometimes introducing new problems, sometimes enabling new and better solutions to existing problems, and sometimes introducing exciting new opportunities that simply were not on our minds before.

Introduction

Massive Online Collaboration is revolutionizing the way in which content is being produced and consumed worldwide. This is bound to also have significant impacts on how we **translate** content. In this paper, I discuss what some of those impacts might be, and invite the translation community as a whole (translators, tool builders and vendors, clients, researchers, educators) to start thinking about the role they want to play in homesteading this new frontier.

The thoughts exposed in this paper are based on my experience and research on Massive Online Collaboration, wiki in particular (Désilets et al., 2005, 2006 and 2007). They are also grounded in an intimate understanding of translation work practices, which I gained through yet to be published ethnographic research. The later is part of a project called OPLT (a French acronym which stands for *"Observing the Technological Practices of Language Workers"*) in which we conduct contextual inquiries[1] with professional translators while they carry out their day to day work.

Massive Online Collaboration: The New Frontier

We live in very exciting times. Indeed, it has become somewhat of a cliché to say that the internet's impact on the world is comparable to that of the printing press. Yet, the only reason we perceive this to be a cliché is that it is true. In less than 15 years, the web alone has given us a world where anyone in industrialized nations has all the information and knowledge he will ever

[1] Contextual Inquiry: http://en.wikipedia.org/w/index.php?title=Contextual_inquiry&oldid=190730351

need right at his fingertips. With projects like the One Laptop per Child initiative[2], this may soon be true for citizens of the third world as well.

As if this was not exciting enough, the internet is also enabling worldwide human collaboration at a scale never possible before. For over two decades, it has allowed loosely coordinated groups of volunteer programmers worldwide to create Free Open Source Software gems like Linux, Apache and Firefox. As unlikely as it might have seemed twenty years ago, these products now compete favourably with software produced by commercial giants like Microsoft. More recently, Web 2.0 technology has allowed a distributed and loosely coordinated community of thousands of authors to collaboratively create Wikipedia, the world's largest online encyclopaedia, and one whose content competes favourably with that of revered sources like Encyclopaedia Britannica (Giles, 2005). Other types of content created or managed through this kind of Massive Online Collaboration include everything from news articles (WikiNews[3]), text books (WikiBooks[4]) movie reviews (IMDB[5]), video (YouTube[6]), photos (Flickr[7]), music preferences (LastFm[8]), social networks (FaceBook[9]) and even large virtual 3D worlds (SecondLife[10]).

The quintessential example of Massive Oline Collaboration (MOC) is of course Wikipedia, therefore I would like to illustrate how it works with an anecdote from my recent experience. Last week, I happened to be looking at the Wikipedia page for Gatineau Park, a national park near my home. As I read the content, I noticed a reference to the Keskinada Loppet, an annual ski race held there each year. Because I live right next to this park and usually participate in that event, I happened to know that its name had just been changed to Gatineau Loppet. So I clicked on the **Edit** link, changed the page content to reflect the race's new name, clicked **Save Page**, and my update was immediately there for everyone else to see. As a result, out of date information on the Gatineau Park was corrected in a matter of a single week!

Whenever I do something like this on Wikipedia or any other wiki site, I feel tremendous power at my fingertips. Of course, those of us who are technically inclined know that there is no rocket science involved and that under the hood, Wikipedia is nothing more than a collection of PHP scripts connected to an SQL database. But that is exactly what is exciting about Wikipedia (and wikis in general): that something this technologically simple could enable complex collective behaviour and lead to the creation of something as large, rich and complex as Wikipedia.

Indeed, consider the following surprising facts about this anecdote:

- The creators of Wikipedia do not know me and I do not know them.
- I do not know the people who created or worked on the Gatineau Park page and they do not know me.
- The Wikipedia software does not know about me since I was not required to login.
- I am but one of thousands of industrious bees who participate in the creation of something big and important.
- I did this without any economic incentive, simply because I care deeply about Gatineau Park.

Yet, it works. Wikipedia is a proof by construction that people are able to collaborate very efficiently and create high quality valuable content in that way.

---

[2] One Laptop Per Child project: http://laptop.org/.
[3] WikiNews: www.wikinews.org
[4] WikiBooks: www.wikibooks.org
[5] Internet Movie Database: www.imdb.com
[6] YouTube: www.youtube.com
[7] Flickr: www.flickr.com
[8] LastFm: www.lastfm.com
[9] Facebook: www.facebook.com
[10] SecondLife: www.secondlife.com

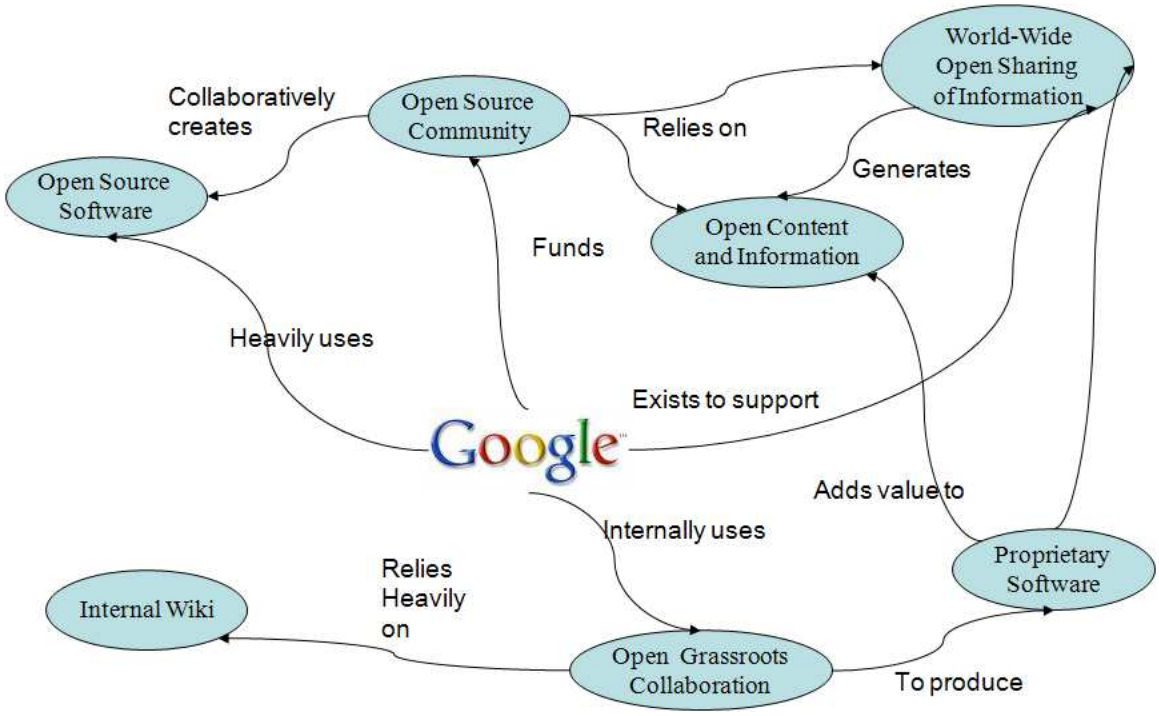


Figure 1: Complex collaboration ecosystem of a modern software company.

Is this the Future of Content Creation?

Wikipedia and other Massive Online Collaboration sites will shake the world of content creation to its very foundation (Tapscott and Williams, 2006, Brafman and Beckstrom, 2006). Indeed, it is quite plausible that in a near future, a significant portion of the content we consume will be created using this kind of powerful new paradigm.

Some may find this statement naïve and idealistic. After all, Massive Online Collaboration may be nice and fine for amateurs who have nothing better to do than write encyclopaedia articles for free. But surely, it is not relevant for professional writers and translators who typically work for **proprietary** content vendors, is it?

Yet, one should be careful about clinging to this sort of assumption, because this is exactly what the software industry used to say about the Open Source movement[11] when it first started being popular in the early 90s. Yet today, many of the most successful software companies are organizations that have learned to thrive in a complex ecosystem of open collaboration which includes Open Source.

As illustrated in Figure 1, a company like Google for example, exists solely for the purpose of facilitating collaboration and exchange of information between people worldwide. In order to fulfill that mission, Google makes heavy use of Open Source software like Linux and Apache, which was built largely through online collaboration. The company recognises this dependence and heavily funds Open Source projects through initiatives like the Google Summer of Code. Of course, Google is not an Open Source company by any stretch of the imagination, and it develops a lot of proprietary software which it keeps close to its chest. But even when doing so,

[11] Open Source Software:
http://en.wikipedia.org/w/index.php?title=Open_source_software&oldid=190779548

Google internally uses processes that are closer to the type of open collaboration found in Open Source, than the traditional top-down command and control paradigms typically found in closed proprietary environments[12]. In particular, it is worth noting that Google uses a very large internal wiki to foster grassroot collaboration and information sharing across the organization (Buffa, 2006).

I suspect similar changes will happen with proprietary content creation. In the future, successful content producers will be the ones who learn to thrive in this world of open, Massive Online Collaboration and who discover ways to either facilitate the production of MOC content, add value to it, or use MOC internally to produce proprietary content. This view is echoed by many thought leaders in collaboration technology (Tapscott and Williams, 2006, Brafman and Beckstrom, 2006).

Is this the Future of Content Translation?

Of course, such drastic changes in the way that content is produced will also change the way in which content is **translated**, and this is the theme of the present paper.

When venturing in this new frontier, there are a number of questions that naturally come to mind with respect to translation. Below is a very partial list.

- What processes and tools are needed to translate content that is constantly being edited collaboratively by a large, loosely coordinated community of authors?
- How might translators benefit from open, wiki-like translation resources?
- Could collaborative semantic tagging help improve Machine Translation by allowing large numbers of people to teach machines facts about the world?
- In a world where anyone can write and publish original content in their native language, will we need to cover more language pairs, and if so, how might Machine Translation technology help?
- Could massively collaborative technologies give freelance translators the kind of competitive edge that technology has more traditionally provided to larger organizations, by allowing them to work within the framework of a large international community of practice?
- Will we see the emergence of a new breed of amateur volunteer translators and will this result in de-skilling of the translation profession and lowering of quality standards for translation?
- How can organizations best leverage the collaborative energy of this new breed of amateur volunteer translators?
- How do we ensure the quality of translations and translation resources in a seemingly chaotic massively collaborative environment?
- How can massively collaborative technologies help linguistic minorities?
- How can massively collaborative technologies help save small languages (e.g. aboriginal languages) from extinction?

[12] For an account by a Google employee, see: http://steve-yegge.blogspot.com/2006/09/good-agile-bad-agile_27.html

- Can teachers of translation take advantage of massively collaborative online environments to provide students with real-life translation experience early on in their training?

- What can the history of Open Source tell us about the future of translation in this new world of open, Massive Online Collaboration?

In the present paper, we will focus on the first three of those questions, because they illustrate the different ways in which Massive Online Collaboration is changing the rules of the game for translation.

On the one hand, MOC may introduce new problems and challenges:

> *What processes and tools are needed to translate content that is constantly being edited collaboratively by a large, loosely coordinated community of authors?*

On the other hand, MOC may also enable new and better solutions to old problems:

> *How might translators benefit from open, wiki-like translation resources*?

Finally, MOC may introduce exciting new opportunities that were simply not on our minds before:

> *Could collaborative semantic tagging help improve Machine Translation by allowing large numbers of people to teach machines facts about the world?*

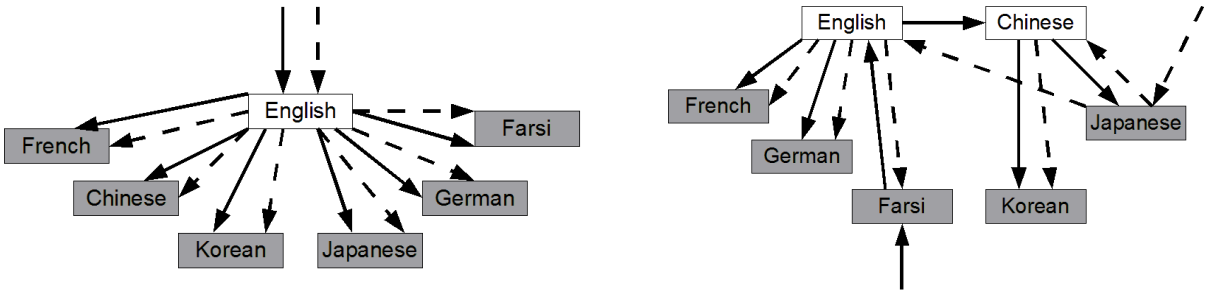
What processes and tools are needed to translate content that is constantly being edited collaboratively by a large, loosely coordinated community of authors?

A central difference between the old world and this new frontier of Massive Online Collaboration is that in the new world, content is not mandated… it just "happens". Indeed, in this type of environment it is difficult to force or even nudge people to do any particular task, since contributors are typically volunteers who are motivated by their own personal interest in a particular content or community (Forte and Bruckman, 2005). Moreover, the governance rules and structure are mostly determined by the community in a collaborative decentralised way (Forte and Bruckman, 2008).

Consequently, many of the traditional top-down, command and control translation paradigms we use today fall apart in a MOC situation (Désilets et al, 2006). For example, Figure 2 illustrates how the translation workflow is much more open and chaotic in a MOC environment than in a traditional environment.

In a traditional environment (Figure 2a), when a document is originally created, the organization can mandate that it be first created in some master language, typically English. Only when the document reaches a stable state in that master language, will it be translated to all the other languages. Similarly for subsequent modifications, the organization can mandate that the change first be done in the master language, and then translated to other languages.

In contrast, one does not have that sort of luxury in a MOC environment (Figure 2b), because it is simply not reasonable to require that all volunteer contributors create original content and modifications in say, English. In fact, even if one could force all contributors to write in English, this might not be advisable since many of them might not be fluent enough to write high quality content in that language. Therefore, a contributor whose native language is say, Farsi, should be



(a) Traditional **(b) Massive Online Collaboration**

Figure 2: Translation workflows in Traditional versus Massive Online Collaboration environments. Solid arrows represent initial document creation while dashed arrows represent subsequent modifications.

allowed to create original content in that language, and the process and tools should facilitate propagation to other languages. To make matters worse, when another contributor changes this content, he may not be in a position to do it in neither English nor Farsi. This more open and chaotic nature of the MOC workflow means that the processes and tools need to support translation along more complex paths, and possibly involving translation between more language pairs than in a traditional context.

Another consequence of the *"content happens"* principle is that in a MOC environment, it is very difficult to ensure timely translation of content. The elapsed time between the moment where a document or modification is first created in one language, and the moment where this document or modification has been propagated to all other languages, can be much longer than in a traditional environment. This means that you cannot realistically wait for new content to be propagated to all other languages before you publish. This in turn means that your processes and tools must be able to deal with situations where substantial parts of the content have only been translated in certain languages.

These are only two of the many challenges that we face when trying to manage and facilitate collaborative translation of content that is produced through a Massive Online Collaboration paradigm (for a more detailed analysis of this question, see Désilets et al, 2006). Yet, researchers and practitioners worldwide have started experimenting with these ideas in real production settings. For example, Wikipedia has a translation project[13] which employs lightweight tools and processes to coordinate a community of volunteer translators. The Cross Lingual Wiki Engine project[14] aims at supporting collaborative translation more explicitly inside popular wiki engines like TikiWiki, MediaWiki and TWiki, using designs and principles described in Désilets et al, 2006. WorldWide Lexicon[15] is a service trying to enable this kind of massively collaborative translation for any web site (not just wiki sites). Other projects are looking at massively collaborative translation of content that is more static and may have been authored in a non-collaborative fashion. For example, TraduWiki[16] is a site for the collaborative translation of documents licensed under Creative Commons, and DotSub[17] is a similar site for translation of video subtitles.

[13] Wikipedia Translation Project: http://meta.wikimedia.org/wiki/Translation.
[14] Cross Lingual Wiki Engine (CLWE) project: http://www.wiki-translation.com/Cross+Lingual+Wiki+Engine+Project
[15] WorldWide Lexicon: www.worldwidelexicon.org.
[16] TraduWiki: www.traduwiki.org.
[17] DotSub: www.dotsub.com.

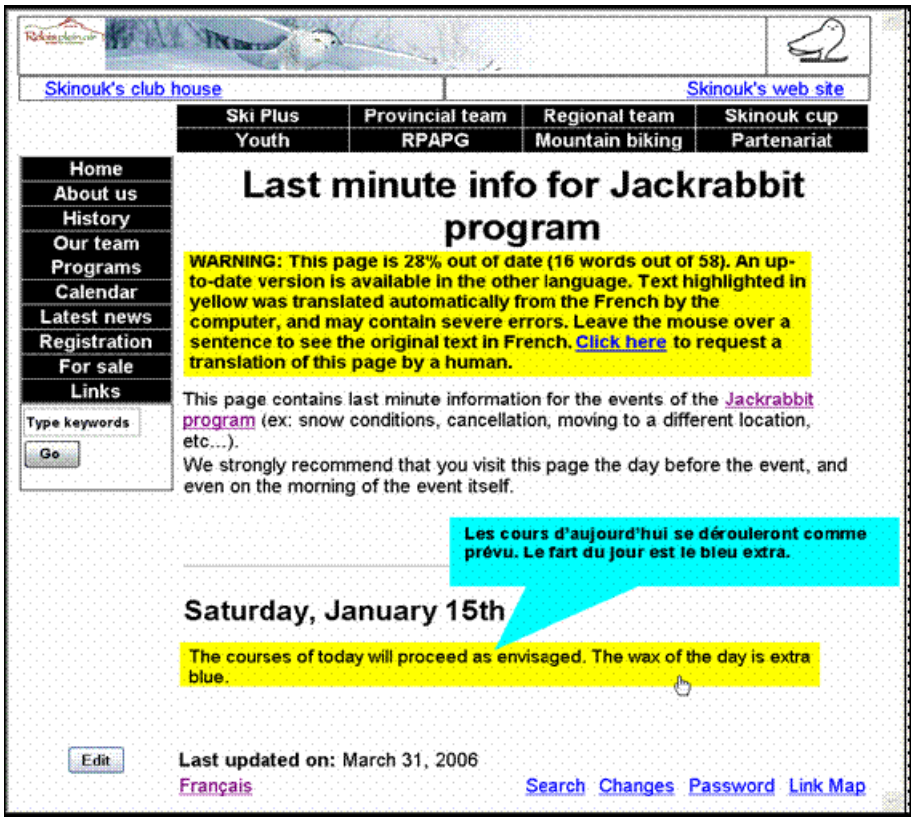


Figure 3: Helping site visitors stay abreast of changes made in other languages.

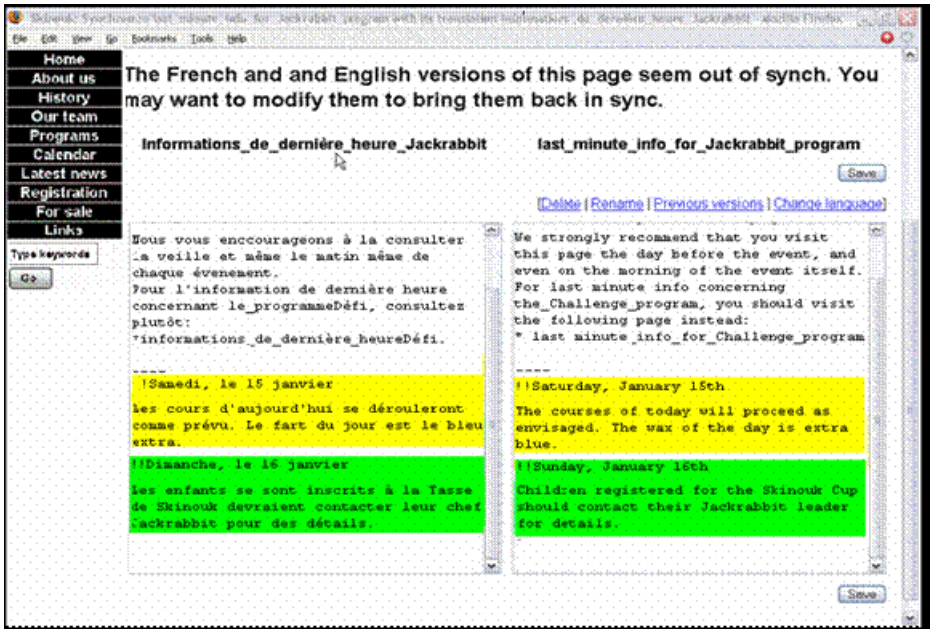


Figure 4: Helping site translators reproduce changes from one language to another.

The limited space of this paper does not allow me to talk about all of those projects in detail. However, I would like to give a flavour of what is being developed today, by discussing some examples taken from the Cross Lingual Wiki Engine project.

Figure 3 shows a page as it might be seen by a site visitor who is interested in reading content in English. It looks very much like a page on any wiki or web site, except that there is a warning at the top telling the visitor that the page is 28% out of date and that a more recent version is available in French. If the visitor happens to be fluent in French, he may then opt to read the page in that language by clicking on the French link provided by the system. On the other hand, the visitor may opt to read the out of date English page instead, in which case the system still helps him stay up to date by providing an automatically generated machine translation of the changes that have been done on the French side, but not yet translated by a human. Portions that have been translated automatically are clearly highlighted so that the visitor knows to treat them with caution. Also, if the visitor has any level of fluency in French she can hover the mouse over the machine translated English text and see the original French text, which may help to clarify any mistranslations done by the machine. We believe that this kind of approach may allow a massively collaborative site to publish original content or modifications made in one language, without having to wait until it has been human translated in all other languages.

Figure 4 shows another screenshot, this one targeted at a site translator who is trying to bring a French and English pages in synch with each other. Here, the system displays the French version on the left and English on the right. Untranslated changes from the French side are highlighted in yellow, while untranslated changes from the English side are highlighted in green. In both cases, the system has automatically pasted a machine translation of the change into the other language and the human translator can use that as a starting point. Note again how only small portions of clearly identified text are automatically translated, while the bulk of the page consists of text that has been already translated or vetted by a human. We believe that this sort of dialog could allow volunteer translators to more easily propagate changes from any language (including languages that they do not understand) to their native language. This in turn means that the site can support a more flexible translation workflow which does not assume that original content or modifications are always made in a master language like English.

These are only two examples of features that can be built to support collaboration of content produced through massive online collaboration, and they are largely untested as of this writing. But I believe that some form of massively collaborative translation tools along those lines will eventually emerge and be successful.

How might translators benefit from open, wiki-like translation resources?

Of course, Massive Online Collaboration does not only introduce new challenges and problems. It may also enable new and better solutions to old problems. In particular, it could lead to the improvement of Terminology Databases (TD) and Translation Memories (TM).

TDs and TMs are the two technological pillars of modern translation. Together, they have resulted in significant gains in translator productivity, and have enabled more consistent translation of large documents by teams. In a way, they are collaboration technologies because they act as repositories of collective expertise. But they are **not massively collaborative** by any stretch of the imagination**,** because the number of people who can contribute content to them is very limited. For example, although a public terminology database like TERMIUM[18] may be consulted by thousands of people worldwide, its content can only be edited by a very small number of highly specialized terminologists at the Translation Bureau of Canada. As for translation memories, they

[18] TERMIUM: www.termiumplus.gc.ca.

have so far mostly been deployed internally inside organizations and they can typically only be consulted and edited by people working for that organization. Moreover, while the TM might be consulted by hundreds of translators (in the case of a large organization), the ability to **add content** to its database tends to be restricted to a much smaller group of experts.

When thinking about the future of translation resources in a massively collaborative world, it is therefore natural to ask what might happen if we were to open up TDs and TMs for consultation **and editing** by a very large community of terminologists, translators, and possibly even members of the public at large.

One possible positive outcome would be to improve the coverage of these resources. In our interviews with translators, we observed that they tend to use a wide range of resources when trying to resolve translation problems. Indeed, it is not uncommon for a translator to have two or three dictionaries and terminology databases opened at the same time, as well as a translation memory and a Google window for searching in corpuses. Translators told us that they find switching between those different resources annoying and counterproductive, and some went through great length to arrange them in their task bar and web browser, so as to facilitate their navigation. Yet, there were many cases where the translator did not find any relevant solution, even after consulting all the resources he had at his disposal. Very large TDs and TMs created collaboratively by thousands of translators worldwide might alleviate these problems, by offering translators large, versatile resources that cover most of the translation difficulties they encounter.

Another potential positive outcome would be to allow freelance translators to share terminology and translation archives with a worldwide community of practice, thus achieving some of the economies of scale which at the moment, only large translation organizations can achieve.

There are currently several projects which can legitimately claim to be already acting as massively collaborative TDs. For example, ProZ[19] is a large online community of translators which supports various modes of collaboration, including sharing of glossaries. Other examples include wiki sites like Wikipedia and Wiktionary[20] (a wiki-based dictionary) which, while technically not TDs, include inter-lingual links between words and terms in different languages. Another example is OmegaWiki[21], a proper multilingual terminology database based on a wiki paradigm. At present, the usefulness of those various wiki resources for translators is somewhat mixed. For example, we have conducted an evaluation of Wikipedia, Wiktionary and OmegaWiki and found that in their current state, they are not very useful to translators. Reasons for this are that they lack sufficient coverage of typical translation difficulties, and that their user interface is not designed for rapid consultation and editing by translators (Désilets et al, 2007). On the other hand, while we have not yet conducted an equivalent evaluation of ProZ, the fact that it is a well known resource in translation circles hints that it might already be useful to translators in its current state. In any case, the issues of coverage and user interface which we have uncovered in our study of wiki resources are far from insurmountable, and therefore it is plausible that in a near future, massively collaborative TDs will become a viable option for translators.

Another way in which massively collaborative TDs might become a reality would be for providers of large proprietary TDs like TERMIUM, GDT[22] and IATE[23] to open up a version of their resources to editing by a wider community. Although this would have been unthinkable even a few years ago, the recent success of Wikipedia might inspire publishers of these resources to follow suite and experiment with a massively collaborative paradigm.

---

[19] ProZ shared lexicons: http://ksearch.proz.com/search/.
[20] Wiktionary: www.wiktionary.org.
[21] OmegaWiki: www.omegawiki.org.
[22] GDT: www.granddictionnaire.com.
[23] IATE: http://en.wikipedia.org/w/index.php?title=Inter-Active_Terminology_for_Europe&oldid=183674051.

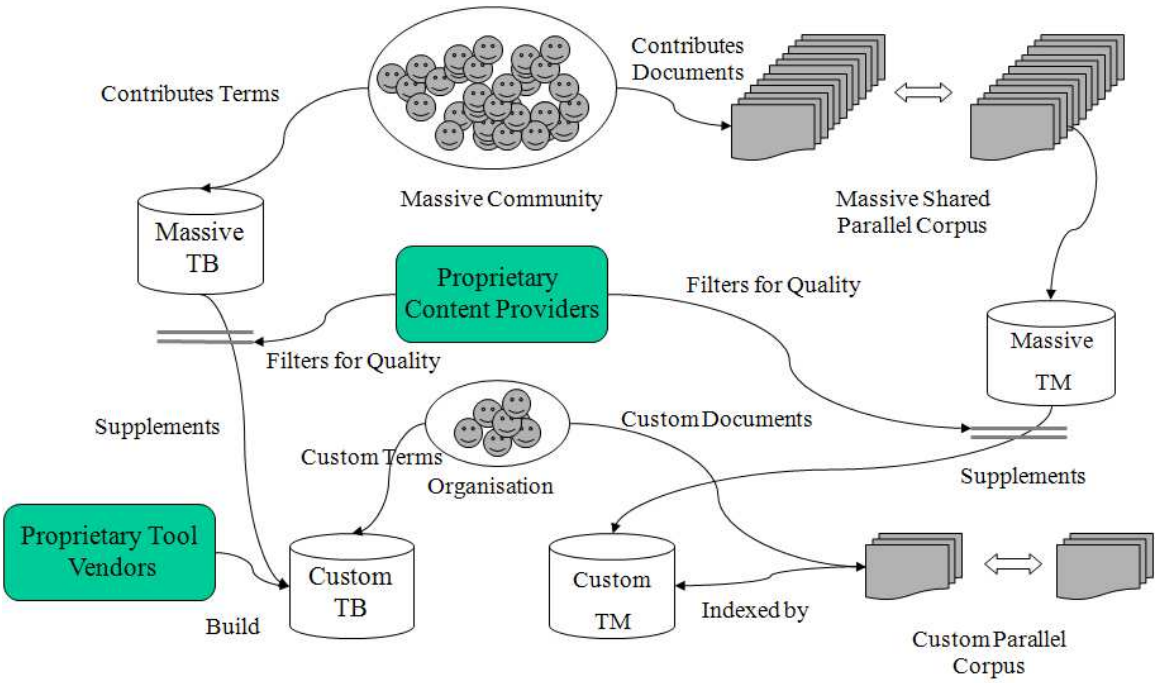


Figure 5: Synergies between free, open MOC tools and content, and closed proprietary ones.

On the TM front, there are also several projects that can legitimately claim to be massively collaborative. For example, VLTM[24] and MyMemory[25] are two projects that aim at building very large translation memories based on parallel texts submitted by wide communities of contributors. In a similar vein, we have started a project called WeBiText[26] to investigate the use of very large translation memories based on parallel content mined from the web. The later may not immediately come across as massively collaborative, until you realize that the world wide web itself is, in a sense, a very large corpus built in a massively collaborative fashion by the sum of all the people and organizations that publish web pages.

Our own experience with the WeBiText project indicates that there is definite value in TMs based on such massive, open and collaboratively built parallel corpuses. For example, we have found that 76% of the English to French translation difficulties observed in our contextual inquiries with professional translators can be answered by a TM based on a million pairs of pages mined from Government of Canada sites. We also evaluated the WeBiText approach on a sample of translation difficulties for the English to Inuktitut pair (Inuktitut being the language of the Inuit people of Canada) and found that 60% of them could be answered by a TM based on 4,300 pairs of pages mined from the Nunavut domain. The later finding is particularly interesting given that current lexical resources for English-Inuktitut are very sparse and do not offer the kind of comprehensive coverage typical of more mainstream language pairs like English-French.

In short, open, massively collaborative TDs and TMs will likely play an important role in the translation industry in coming years. While it may seem naïve to think that one can arrive at high quality linguistic resources through this kind of process, one must remember that just five years ago, the idea of Wikipedia also sounded very naïve to most people. Yet, we now have a proof by construction that it can work. If there is one thing that can be learned from the success of

[24] VLTM: http://www.wordfast.net/?whichpage=jobs.
[25] MyMemory: http://mymemory.translated.net/.
[26] WeBiText: www.webitext.com.

Wikipedia, it is this: **when dealing with collaboration at that kind of scale, our intuitions about what can and cannot happen are often wrong**.

Will the availability of free and open TDs and TMs mean the end of proprietary tools and content? I do not believe so. After all, the emergence of Open Source has not meant the end for vendors of proprietary software. In fact, some vendors like IBM developed highly lucrative business models based on Open Source. Likewise, I believe there will still be lucrative niches for proprietary translation tools and content, but vendors will need to find ways to act synergistically with open content. Figure 5 provides an illustration of how that might work.

On one hand, we will have very large communities of terminologists and translators collaboratively sharing content on very large open resources. These resources will form a solid free, open backbone that everyone can benefit from. On the other hand, we will also have a large number of organizations, each translating in particular domains and for specific purposes, and with particular workflows and operational platforms. This is where proprietary tool vendors will be able to act, by providing such organizations with tools that are perfectly suited to their needs and can work seamlessly across the organization's workflow. Tool vendors may also be able to sell services and support to help organizations deploy such tools effectively. These proprietary tools will however need to link with larger open MOC resources if they are to provide optimal value to users and organizations. For example, if a particular term or expression is not found in the local proprietary resource, the system might then look on the larger but more generic open resource.

Organizations that currently produce proprietary content like the TERMIUM and IATE databases, may also have a role to play in this open world. For example, they may be able to sell content that is complementary to open resources, such as Seals of Approvals that help users distinguish between high quality and poor quality content.

Could collaborative semantic tagging help improve Machine Translation by allowing large numbers of people to teach machines facts about the world?

Up to now, I have mostly talked about things for which the writing seems to be on the wall. The question is not whether or not they will happen, but rather when and how, and what level of impact they will actually have on the world of translation.

With my third question, I am now leaving that comfort zone, and entering a world of speculation. But I offer it as an example of something truly groundbreaking that massive online collaboration might achieve for translation technology.

Any translator will tell you that it is very difficult to translate text that one does not understand. Yet, this is exactly the situation that Machine Translation (MT) technology is in today. Although MT systems may know quite a lot about language (either in the form of hand-crafted rules or statistical patterns learned from large corpuses), they know absolutely nothing about the world that we talk about through language.

It is therefore natural to ask whether the accuracy of MT systems could be improved by providing the machine with world knowledge. The answer to this question is far from obvious. The idea of grounding MT in explicit world knowledge has been around for a while and was relatively popular in the late 80s to early 90s (for example, see Knight and Luk, 1994). Although this knowledge based approach enabled development of high quality, fully automatic MT systems in small restricted domains like weather forecasting (ex: METEO[27]), it does not scale up to large general

[27] METEO: http://en.wikipedia.org/w/index.php?title=METEO_System&oldid=177763004

domains, because manually writing semantic translation rules quickly becomes an unmanageable burden. In contrast, modern Statistical MT systems have been able to achieve reasonable quality levels on very large unconstrained domains, without having to explicitly represent semantic information about the world (Och, 2005). However, some researchers feel that this type of brute-force statistical approach is reaching a plateau which can only be surmounted by injecting more explicit human knowledge into the algorithms (Rosenfeld, 2000). Indeed, in the related field of text classification, some researchers have recently found that major "disruptive" gains in accuracy can be achieved by augmenting statistical methods with human knowledge embodied in Wikipedia pages (Gabrilovich and Markovitch, 2005).

Unfortunately, building a machine that knows about the world and can use that knowledge intelligently turns out to be extremely difficult. This problem, which Artificial Intelligence (AI) researchers refer to as Common Sense Reasoning, is one of the holy grails of AI. But contrarily to other AI problems like speech recognition and chess-playing, it has so far resisted all attempts at solving it. Surprisingly enough, creating a machine that can beat Kasparov at chess (Hsu, F. H., 1999) turns out to be much easier than building a machine that knows as much about the world as a five year old!

The most dramatic illustration of this difficulty is the CYC project (Lenat, D., 1995), which aims at creating a manually built database of common sense facts about the world. The project, which started in 1984, has been going for more than two decades, and is reported to have cost in the order of $25 million. Yet, there is no evidence that it is achieving the kind of major impact one would expect from a fully functioning Common Sense Reasoning system.

Massive Online Collaboration might however change the rules of that game. Indeed, the reason why technologies like CYC and Knowledge Based Machine Translation failed is that it is simply not possible for a few hundreds of people to write down everything that a machine might need to know about the world in order to carry out an open ended task like translating text in unconstrained general domains. But what if millions of people each wrote down a small number of machine-readable facts about those parts of the world that they care and know most about, and shared that data on a central wikipedia-like resource? This is an idea that is gaining a certain amount of momentum in some research circles, as evidence by ideas and projects like the Semantic Web[28], SemanticWiki[29] and Open Mind Common Sense[30]. In particular, SemanticWikis are wiki sites that provide authors with fast and easy ways to tag parts of texts and turn them into machine readable facts. For example, someone editing a wiki page about the city of Ottawa might write the following text:

> *Ottawa is the capital of [[Is capital of::Canada]]*

which explicitly codifies a particular relationship between Ottawa and Canada in a machine readable form. At the moment, very few wiki sites support this sort of tagging, and even on those that do, few authors go through the trouble of tagging text that way. But what if this became as common practice as bolding or italicizing text? At the moment, this seems unlikely to happen because semantic tagging is still an awkward process. Moreover, there is a chicken-and-egg issue. Users will only feel compelled to write this sort of tags if there are useful applications that can leverage them, yet, useful applications cannot be built until there are enough semantic tags on the web to reach critical mass.

However, as the Wikipedia experience shows, critical mass can sometimes be reached very rapidly in a massively collaborative environment, especially with appropriate technology for facilitating collaboration. In the case of semantic tagging, an appropriate technology might be a

[28] Semantic Web: http://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=179181797.
[29] Semantic Wiki: http://en.wikipedia.org/w/index.php?title=Semantic_wiki&oldid=172351981.
[30] Open Mind Common Sense: http://commonsense.media.mit.edu/.

system that can learn patterns from human generated tags, and then use those patterns to automatically process text that has not yet been tagged by humans (for a survey of this type of technology, see Gomez-Pérez et al, 2003). Those automatically generated tags could in turn be vetted or corrected by humans, and become additional examples from which the machine can learn.

One indication we have that collaborative semantic tagging might help with Common Sense Reasoning, is the fact that social tags have been used successfully to facilitate Content-Based Rich Media Retrieval. This too is a difficult AI problem where a machine is asked to find Rich Media files (image, audio, video), based on a free form query about its content (for a recent survey, see Lew et al, 2006). For example, a query might be *"find a photo of a tree growing in the desert"* or *"find a melancholic song"*. While machines may easily recognize words like "tree" and "melancholic" in text, they currently have no reliable way of assessing whether a photo contains a tree, or what mood a song might induce on human beings. Yet, if you go to sites like Flickr[31] and LastFm[32], you can enter queries of this sort and find relevant Rich Media. The "magic" behind those services is very simple. They have harnessed the collective energy of thousands of users, by offering them simple tools that allow them to easily tag images and songs with words that describe the content from their own personal point of view.

The fact that social tagging was successful in cracking the difficult problem of Content-Based Rich Media Retrieval tells us that massively collaborative knowledge bases might also make a dent on the even more difficult problem of Common Sense Reasoning. If so, this might turn out to be highly useful to help Machine Translation systems do better by allowing them to actually "understand" the meaning of what they are translating.

Conclusions

In summary, Massive Online Collaboration is revolutionizing the way in which content is produced and consumed worldwide, and this is bound to also have a large impact on the way in which content is translated.

On the one hand, MOC introduces new challenges and problems, such as dealing with translation of content created through open, collaborative, and somewhat chaotic workflows.

On the other hand, MOC may also enable new and better solutions to existing problems, for example, by allowing massive communities of translators and terminologists to collaboratively build very large Terminology Databases and Translation Memories.

Finally, MOC may open up brand new opportunities, such as the possibility of improving Machine Translation through the use of large world knowledge bases built in a massively collaborative fashion.

Members of the translation world should all be thinking about the role they want to play in homesteading this new frontier, whether they be translators, clients, tool builders and vendors, educators or researchers. A good place to discuss these issues is wiki-translation (www.wiki-translation.com), an online community started in November of 2007, specifically to discuss and share knowledge about Massive Online Collaboration in the world of translation. In thinking about these issues, one must be careful about assuming that such and such thing cannot happen, because the "laws" that govern collaboration at this kind of scale are not yet well understood, and they appear to be very different from the laws that govern collaboration and social interaction in our day-to-day experience.

[31] Flickr: www.flickr.com.
[32] LastFM: www.lastfm.com.

Acknowledgements

This paper is the result of online and corridor discussion with a number of very bright people, including (but not limited to): Louise Brunette (UQO), Christiane Melançon (UQO), Jean Quirion (UQO), Geneviève Pateneaude (UQO), Marta Stojanovic (NRC), Benoit Farley (NRC), Caroline Barrière (NRC), George Foster (NRC), Pierre Isabelle (NRC), Roland Kuhn (NRC), Peter Turney (NRC), Marc Laporte (TikiWiki), Gerard Meijssen (OmegaWiki), Kizu Naoko (Wikipedia), S.J. Klein (MIT Media Lab), Louis-Philippe Huberdeau (ETS), Xavier de Pedro Puente (TikiWiki), Jérémie Leblanc (NRCan), Peter Cowan (NRCan), Colas Nahaboo (ILOG), Sébastien Paquet (UQAM) and Lucas Gonzalez (Fluwiki).

References

Brafman, O., Beckstrom (2006) R. A. *"The Starfish and the Spider: The Unstoppable Power of Leaderless Organizations".* ISBN, 2006.

Buffa, M (2006). "*Intranet Wikis".* Intraweb workshop, WWW Conference 2006, Edinburgh.

Désilets, A., Barrière, C., Quirion, J (2007). *"Making Wikimedia Resources more Useful for Translators".* WikiMania 2007, The International Wikimedia Conference, Taipei, Taiwan , August 3-5, 2007.

Désilets, A., Gonzalez, L., Paquet, S., Stojanovic, M (2006). *"Translation the Wiki Way".* Proceedings of WikiSym 2006 - The 2006 International Symposium on Wikis. Odense, Denmark. August 21-23, 2006. NRC 48736.

Désilets et al, Paquet, S., Vinson, N.G. (2005). *"Are Wikis Usable?".* Proceedings of WikiSym 2006 - The 2005 International Symposium on Wikis. San Diego, California, USA. October 17-18, 2005.

Forte, Andrea and Amy Bruckman (2008). *"Scaling consensus: increasing decentralization in Wikipedia governance"*. To appear in the Proceedings of Hawaiian International Conference of Systems Sciences (HICSS).

Forte, Andrea and Amy Bruckman (2005). *"Why do people write for Wikipedia? Incentives to contribute to open-content publishing"*. GROUP 05 workshop: Sustaining community: The role and design of incentive mechanisms in online systems. Sanibel Island, FL.

Gabrilovich, E., Markovitch, S (2006). *"Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge"*. In Proceedings of the Twenty-First National Conference on Artificial Intelligence, pages 1301-1306, Boston, MA, 2006.

Giles, J. (2005). *"Special Report: Internet encyclopaedias go head to head".* Nature 438, pp. 900-901, (15 December 2005).

Gomez-Pérez, A., Macho, D., M., Alfonseca, I., Nez, E.N., Blascoe, I., Staab, S., Corco, O., Ding, Y., Paralic, J, Troncy, R (2003). *"Ontoweb deliverable 1.5: A survey of ontology learning methods and techniques"*.

Hsu, F. H. (1999). *"IBM's Deep Blue Chess grandmaster chips"*, IEEE, 1999, IEEE Micro, Volume 19 , Issue 2 (March 1999).

Knight, K., Luk, S. K. (1994). *"Building a large-scale knowledge base for machine translation".* Proceedings of the Twelfth National Conference on Artificial intelligence (vol. 1), Seattle, Washington, United States, Pages: 773 – 778, 1994.

Lenat, D. (1995). "*CYC: a large-scale investment in knowledge infrastructure",* Communications of the ACM, Volume 38 , Issue 11 (November 1995).

Lew, M., Sebe, N., Djeraba, C., Jaine, R. (2006). ***"**Content-based Multimedia Information Retrieval: State of the Art and Challenges".* ACM Transactions on Multimedia Computing, Communications, and Applications, pp. 1-19, 2006.

Och, F. J. (2005). *"Statistical Machine Translation : Foundations and Recent Advances"*, (Tutorial). MT Summit 2005, Phuket, Thailand. http://www.mt-archive.info/MTS-2005-Och.pdf.

Rosenfeld, R. (2000). *"Two decades of statistical language modeling: where do we go from here?".* IEEE Proceedings, Special Issue on Spoken Language Processing, vol. 88, n^{o} 8, pp. 1270-1278, 2000.

Tapscott, D., Williams, A. (2007). *"Wikinomics: How Mass Collaboration Changes Everything.".* ISBN: 1-59184-138-0.