GeekSpeak

What's Up?

Jost Zetzsche
jzetzsche@internationalwriters.com

The GeekSpeak column has two goals: to inform the community about technological advances and at the same time encourage the use and appreciation of technology among translation professionals. Jost also publishes a free technical newsletter for translators (www.internationalwriters.com/toolkit).

(Note: An earlier version of this article appeared in the Translation Journal, *http://accurapid.com/Journal.)*

While this article will appear in 2010, I am writing it at the tail end of 2009, a time when newspapers and news sites are full of the 10 bests and 10 worsts of the year. I am not sure that I am interested in the "bests" and "worsts" in translation. (The "funniest"—think Clinton's "Reset" faux pas—would be more enticing.) Instead, what I really am interested in and concerned about is the kind of new developments we have seen in translation in the past year or so and what will be presented to us in the next.

One thing that comes to mind immediately has to be *crowdsourcing*. I have written about this in *The ATA Chronicle* in the past and have tried to encourage us not to have the knee-jerk response that so many of us first had when we rejected the encroachment of crowdsourcing into our territory, a territory where we felt a sense of entitlement. I have encouraged the translation industry to be actively involved in shaping this (only seemingly) new concept into an opportunity that we can live with and profit from.

Another of 2009's notable topics was clearly *machine translation.* If you have not been involved in several discussions about machine translation with colleagues or other peers this past year, it is time for you to go out and get a social life! Companies like Google, Microsoft, Asia Online, and others have been pushing us to reconsider the applicability of machine translation on the basis of usability. The very concept of quality—which we also have had a love-hate relationship with for a long time—has seriously come under fire. The argument goes that since translation quality is very abstract and arguable (yes, we would all agree here), the only relevant measure for translation is usefulness. For some kinds of texts, high stylistic standards are very important (think: literature, marketing); for others it is accuracy (think: legal, medical); and for still others the only thing that counts is the transfer of information (think: social networks, some technical documentation, customer support data). You may disagree with those classifications, but these are the lines that many very large corporations are drawing when deciding what to give to translators and what to have machine translation do.

Then there is another topic that has come charging to the forefront in just the past few months: the *availability of large amounts of bilingual data* that can be used in translation memories (TMs). Here is just a sampling:

MyMemory
mymemory.translated.net

A colossal TM of presently around 300 million segments that contains data from Web alignments (approximately 30% of the total data), corpora such as the EU corpus (approximately 50%, see DGT TM below), and TM contributions from translators. It offers terminology searches, download and upload of translation memories in TMX (the Translation Memory eXchange format), editing capabilities for users, and a strong tie-in to machine translation.

BigTM
www.bigtm.net

A custom translation search engine that can be used by language services providers or translators. You can submit the translatable text or a sample of it, and the system goes out on the Web to search for pages similar to the source text that already have translations in the target language. Within 24 hours it then provides a searchable index of the discovered parallel pages that allows you to look up how terms or phrases were translated by others in the past. (This product is still in its beta phase.)

OPUS
urd.let.rug.nl/tiedeman/OPUS/

An open-source parallel corpus with a large number of bilingual files in many language directions containing such varied material as data from the European Medicines Agency, the European constitution, the European Parliament Proceedings, the OpenOffice .org corpus, the opensubtitles.org corpus, and various open-source localization and software documentation files. The author of the site is a researcher working in natural language processing and machine translation, so the files are not especially made for translation memory—most of them are in a text format—but they are nothing that could not be converted to a TM-compatible format or even TMX (and the files for the European Medicine Agency are in TMX).

TAUS Data Association
www.tausdata.org

The TAUS Data Association (TDA) is an association of mostly large corporate translation buyers who originally came together to pool their TM data to train their machine translation engines better. TDA has now just announced that they have launched a relatively low-priced professional membership category that allows you to download 10 times the amount of data that you upload. Also, as a "by-product" they have opened the data up to the public as a terminology resource. Both the terminology searches and the TMX download can be categorized according to client and a (rather coarse) subject taxonomy. Presently (December 2009), the complete corpus includes about one billion words.

DGT TM
langtech.jrc.it/DGT-TM.html

This is the humongous TM for the Acquis Communautaire (the body of EU law) in 22 languages and a total of 231 language pairs. It is available as a free download and the data is presented in TMX format.

Linguee
www.linguee.com

A very large corpus of English-into-German-into-English data (other language pairings will be released in 2010) of Web-based translated materials. The Web-based data is matched up with the help of a large custom dictionary and other Web-based dictionaries. Though the data is not categorized, every entry is accompanied by a link to the originating webpage where webpages or whole websites can be downloaded and aligned (that is, converted into a TM).

And then there are translation environment tools (TEnTs), like Lingotek's suite of tools, Google Translator Toolkit, and Wordfast's VLTM, that are built around the concept of anonymous data sharing through translation memories or alignment tools like AlignFactory and NoBabel's AutoAligner that have finally made alignment of large amounts of Web-based contents feasible.

So what are we going to do with all of this? Is this sudden flood of data going to be helpful or harmful to our productivity via TM technology? The short answer is: I do not know, but I do have an inkling.

When I first started using TEnTs, I was very eager to build up my own data so that I could benefit from my past labor. My "Big Mama TM" grew and grew, and I was always excited to find matches from (almost) forgotten previous projects. As the years passed, I continued to use and feed my meanwhile obscenely large Big Mama TM, but her usefulness seemed to decline rather than improve. Too much time had passed between the earlier projects and the current ones to really classify them when matches were displayed (despite every translation unit being described with subject and client information). In addition, language had changed, as had my skill level, prompting me to spend a lot of time deleting or wading through useless suggestions from the TM. The fact that many of the newer TEnTs now also offered subsegment matching that allowed them to dig even deeper into the language materials did not help either.

I have increasingly come to realize that while large amounts of data are very powerful, they can also be very distracting if they a) originate from a subject matter or client that uses a different terminology or style, b) come from dated or obsolete sources, and c) come from sources with a different quality level.

So what does it mean to have all these gigantic data vaults at our disposal if my conclusions are true? I think that many of them are fantastic as reference material, but I am just not sure about their value as TM data in the classic sense. And it is important to keep in mind that many of these resources were not produced for TM purposes (even though that may be their origin), but to feed the ever-hungry statistically based machine translation engines with their favorite food: bilingual data.

Am I suddenly advocating the dismissal of TM technology? Not on your life! I still think that TM technology in concert with terminology resources should form the foundation of the toolkit of every translator who works on functional texts. But I have also come to the realization that raw data, including TM data, has no value per se. The value of data for the human translator is in its quality and appropriateness.

That is what's up in my eyes!

ata