GeekSpeak

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).

Get Those Things Out of There!

I'm sorry, but I cannot help myself. Earlier today, I was listening to my favorite interviewer, National Public Radio's Terry Gross, talking to someone who had authored a book on the history of child-birthing in the Western world. One birthing tool that particularly engaged the author's and interviewer's fascination was the forceps. (How I remember that blue, Martian-like creature that was to become my first child in the grip of that contraption!) The author talked about the awkward situations created after the introduction of forceps, when doctors were still not allowed to see the uncovered mother-to-be. A tent-like structure was erected around the woman with the doctor's head outside and his hands, holding the forceps, inside blindly feeling their way. I am going to stop here, but as I was listening with wide eyes to that story, I realized that it is a perfect illustration of translation without proper terminology. (Just in case you do not see the value of this story as an illustration, know that it was still tremendously satisfying to retell it.)

A couple of years ago, I wrote a column for this magazine that I named *Captain Sutter's Story.* It was about the nuggets of gold lying around in the data that we amass in our translation work as we build up translation memories, and in all the bilingual data in other formats we have accessed. One of those often-overlooked nuggets is the terminology contained in that data. While we can access some of it through the concordance feature of translation environment tools (the manual search for terms or phrases within longer segments), it would be much more helpful to have those terms sitting in a terminology database. Enter terminology extraction.

The concept of term extraction is, of course, the ability to extract monolingual or bilingual terminology from document(s) or databases to create glossaries quickly that will aid you with your translation projects. One reason that the termbase functionality is still so crudely underused with most tools is that it is tedious to use. (And though it is actually no longer tedious to use in current versions of tools, it used to be tedious, and the user's mind still classifies it as such.).

So, wouldn't it just be great if we could spend a couple of hours before a large project either harvesting terminology from existing projects of the same subject matter or quickly creating lists of source terms that are relevant to our project and translating those ahead of time? (Of course, this all becomes much more relevant and important when you are faced with multi-translator projects.)

Let's start with the no-brainer solution of extracting bilingual data from existing sets of translated documents or databases (typically in TMX format).

- The most powerful application in the field of term extraction used to be the Xerox Terminology Suite (XTS), which was designed for the deep pockets of corporate users and was very powerful because it was based on preconfigured linguistic data in various languages. Today, the suite is owned by TEMIS, where development has virtually (and literally) come to a halt. However, the translation environment tool Similis has integrated the XTS engine and therefore comes with a very high-level linguistic "knowledge" in seven languages of the European Union (English, Dutch, German, Spanish, Italian, Portuguese, and French). Similis is able to apply a combination of linguistic and statistical rules to a number of processes, including automatic extraction of terms and phrases from translation memory content, with extremely high accuracy—but unfortunately only in a handful of languages.

- SDL offers two separate programs—MultiTerm Extract and SDL PhraseFinder (which are sold as a bundle)—that extract existing terminology and build up terminology databases or glossaries and present you with a proposed translated terminology list. MultiTerm Extract, the tool that originally comes from the Trados side of things, works on a purely mathematical level ("if word A always appears in sentences for which word B always appears in the translated sentence, then these words must form a word pair"). This means it supports all Windows-based languages. PhraseFinder, the former SDLX companion, works on a language-based level for English, French, German, Spanish, Dutch, and Portuguese. This means that overall all languages are supported, but users of the languages that are supported by PhraseFinder have drawn the longer straw since the recognition will be more accurate. (On the other hand, the PhraseFinder process is very resource-intensive and not particularly fond of large amounts of data.)

- MultiCorpora's MultiTrans has always offered the extraction of monolingual term lists. With its latest version it added the WORDAlign feature that internally creates bilingual term lists to improve the accuracy of the alignment, but then can also be extracted as separate termbases.

- Another terminology extraction tool is Terminotix's SynchroTerm.

SynchroTerm theoretically supports all languages; however, practically speaking, there are different tiers of language support. In general, SynchroTerm relies on mathematical calculations to extract terminology pairs. For English, French, Spanish, German, Italian, Portuguese, Swedish, Russian, Greek, Polish, Turkish, Dutch, Hungarian, and Norwegian, it also uses lists of stop words to filter out terminology pairs automatically. For English and French, it also makes use of stemming rules, further improving the accuracy in those languages.

These are the ways to create bilingual glossaries semi-automatically. Of course, if you start with just source documents, there need to be ways to extract just source terminology. You have two choices for that. You could use an integrated feature like those offered by Déjà Vu, Swordfish's Anchovy, and MultiTrans. These features create an index of all terms and phrases in the project and allow you to choose, for example, how long the phrases are to be or how many occurrences they are to have. This is very helpful, and if at first it seems that there is a lot of useless material extracted, it is up to you to find good workflows to locate the good stuff quickly and delete the rest.

There are also external tools. These are called "concordancers." You can find a list of concordancers on Wikipedia or in your search engine. You will quickly see that most of them come from academia—clearly there is an interest for linguists to be able to analyze large corpora of text for the actual usage of terms and phrases. But there is also strong interest for us. Many of these concordancers are language-specific, which means that they come with information on what kind of terms or combinations of terms to ignore, but that makes them only more interesting for us.

It is possible that 150 years from now (or maybe just 15!), translators will look back on today's translation processes as groping in the dark—failing to use the appropriate tools to get to the data that is already there—and experience a feeling of horror similar to what we feel today when we read of those blind forceps.