Too Much Data?

I've been pondering a lot about one of the most dramatic changes we've seen in the world of translation. And now you say: Yes, I know, machine translation! You're right that this is a big change, but it's not really what I've been thinking about (at least not exclusively).

I've been thinking about data. We have access to so much data! While there are clearly some advantages to this, there are plenty of disadvantages as well. Or maybe a primarily one: we need to learn to become (better) data curators. (I can't remember ever having heard of that as a topic of study in a translation program.)

What's this data I'm talking about? Okay, let's start with the most obvious: (y)our termbase and translation memory (TM) data.

For years, I—along with many others—was a proponent of the big mama TM and big papa termbase, which essentially consisted of just one repository each for all your TM and termbase data. We had good reasons. I've had cases where an entire file in a new project was virtually identical to a file in a much earlier project, even though the projects didn't seem to have anything in common with each other. If I hadn't used my big mama TM at that point, I would have spent a few hours of unnecessary translation work. But things have changed since then.

First, if you've been a translator since the early days of TM use, those big mamas have become monstrous mamas, slow and unwieldy. More importantly, the technology has changed. With most tools using some kind of automatic subsegment retrieval, the data that otherwise would likely never have been touched again is being processed all the time, resulting in suggestions that are often unhelpful. The concept of sub-segmentation or fragment reuse is fantastic, as long as you have a TM that reflects what you're currently translating. Otherwise it will slow you down at best, or decrease the quality of your output at worst.

We have access to so much data! While there are clearly some advantages to this, there are plenty of disadvantages as well.

The termbase is different because it's not being used for more granular data access like the TM. But the older big papa gets, the more confusing will be his suggestions. Common termbases that can be used profitably for all kinds of customers are useful for only the very few among us who have such a narrow field of expertise that there really is only one set of terminology, or who have only the one client who always lets us work on the same kind of material. For a freelance translator, however, that is the very rare exception.

Instead, for TMs and termbases we've already had to learn how to silo and curate data. But what about other data?

Large terminology databases such as IATE, the EuroTermBank, or TermCoord are clearly very helpful. Typically well researched and replete with sources, subject matter information, and other metadata, these troves of data have been used by most technical translators (working in the respective language combinations) as long as they've been available. These data sources can become problematic, though, when we bring them into our translation environment tools. I would venture to guess that most translators who have tried to use the EuroTermBank plugin in memoQ or have brought IATIS via TBX export into their translation environment have felt a bit like they were drowning in a sea of data. Yes, this trove might contain the lifesaving term, but only after looking very long and very hard.

Years ago, I found a glossary of maybe 20,000 English>German generic terms. I was so excited to import it into my master termbase, expecting it to greatly increase my productivity. Instead, it made me waste several hours identifying and deleting those terms after a few days of fighting with extreme data overload. Naturally, the data that forms the glossaries and termbases mentioned previously is not generic, but it tends not to be helpful right in your translation environment unless you're able to narrow it down to a much greater detail than is offered.

And then there are the many corpora out there that just seem to be waiting to be turned into TMs. These include the massive DGT TM of the DGT (European Commission), the famous English<>French Hansards corpus of the Canadian Parliament, the United Nations corpus in Arabic, Chinese, English, French, Russian, and Spanish, and the many other corpora that could be used as TMs and are listed on the OPUS corpus page.

Yes, they can all be used. But are they beneficial? They do benefit machine translation (MT) developers as fodder for their engines (provided they are building generic engines), but for translators they are in all likelihood overkill.

Does this mean we should completely ignore these data sources? Probably not, but only if we can find better ways of using them than in the brute force manner of bringing them in as regular TMs. As reference material that we can consult for some tricky phrases they might be great. But that's what the

This 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.

We need to be curators to select the right data from which we will choose, as well as "on-the-fly-curators" to make the right choices in a timely and effective manner as we work.

makers of the Linguee online dictionary figured out long ago, and you can get much of the information via that route as well.

That brings us to MT. In addition to all the (potential) data mentioned above, there are also various MT engines you can consult as you translate. In many cases you can connect not only one but several of those engines within your translation environment at the same time and use their suggestions or parts thereof.

The questions then become obvious: Where is the limit to our capacity to process all that data with which we're bombarded? At what point do our brains go into overdrive and wear us out much earlier in the day than they used to? When do we just choose something to "get it done with" rather than produce high-quality translation?

These are very important questions and I don't know that anyone has the answer. For one thing, I suspect that the answer varies from translator to translator. Also, I think it's a matter of training. Though I'm not aware of any program or course that offers this (though I imagine it might be a very successful endeavor if done right), we'll have to train ourselves not to become distracted by information presented all around the target field where we have to enter the translation. We need to be curators to select the right data from which we will choose, as well as "on-the-fly-curators" to make the right choices in a timely and effective manner as we work.

Martin Kappus, who teaches at the ZHAW Zurich University of Applied Sciences in Switzerland, explored part of this question in a posting on the Language Technology Wiki recently:

> Nowadays machine translation suggestions are dynamically generated and presented to translators and post-editors. They are even adapted depending on the input by the translator/post-editor. It seems that these new methods yield better output from MT and they also seem to get the translator more "involved" in the post-editing process. Do these additional resources pose additional cognitive load on the translator/post-editor? Particularly when working in longer segments where the suggestions change frequently and rapidly?[1]

Should we use the space provided by the Language Technology Wiki (www.langtech.wiki) to discuss all of this? I think it would be a very profitable debate.

SITES MENTIONED IN THIS COLUMN

DGT TM
http://xl8.link/DG-TM

EuroTermBank
www.eurotermbank.com

Hansards Corpus
http://bit.ly/Hansards

IATE
www.iate.europa.eu

Language Technology Wiki
www.langtech.wiki

OPUS
opus.nlpl.eu

TermCoord
http://xl8.link/TermCoord-databases

United Nations Parallel Corpus
http://xl8.link/UNCorpus

NOTES

1 Martin Kappus' posting regarding MT can be found on the Language Technology Wiki at http://bit.ly/Martin-Kappus.

Jost Zetzsche is the author of *Translation Matters*, a collection of 81 essays about translators and translation technology. Contact: jzetzsche@internationalwriters.com.