The Translation Technology Run-down

Jost Zetzsche

Imagine this scenario: A new face shows up at the 2008 regional translators' annual gathering and introduces herself as someone who just got started as a translator. She admits that languages are not her strong point, but says she's sure she can get by with good dictionaries and spell-checkers. I'll predict she'll be spending that party pretty much left to herself.

Now imagine the same scenario with a twist: This time the new translator says that she feels very strong linguistically, but, boy, her computer must have crashed six times yesterday and she can't even install the latest version of Microsoft Office, let alone specialized programs for translators. I predict that she will be surrounded by a couple dozen translators all too eager to chime in with "Me, too! Me, too!"

It's clear that translation professionals come from different stock than, say, engineers. Here is an interesting way to prove that point. When was the last time you went to an engineering website and found an image of the patron saint of engineers, St. Patrick, or the cool patroness Lady Godiva? I'm sure it's been awhile. How about translation websites with St. Jerome, the patron saint of translators? There must be hundreds! And though I see no problem with identifying with one of the giants of our profession's history, it's dangerous to get stuck.

So, given our industry climate, is translation technology an oxymoron? Not on your life! It's just that getting translators to use it is sometimes about as easy as making your kids clean their rooms or brush their teeth.

Translation technology: ready, set, go!

About ten years ago, long-time translation technology veteran Alan Melby released a typology of "Eight Types of Translation Technology" (see www.ttt.org/technology/8types.pdf). They consist of:

- 1: Infrastructure
- 2-4: Term-level before, during and after translation
- 5-7: Segment-level before, during and after translation
- 8: Translation workflow and billing management

Interestingly, the order of the items corresponds loosely to the timeline with which the language industry attached importance to them.

The first principle, infrastructure, is concerned with communication, systems to create and manage documents, and database capabilities. This infrastructure formed the technological basis that allowed us to use translation technology in the first place and turned us from individual service suppliers into a relatively well-connected industry.

St. Jerome — patron saint of translators.
(Peter Paul Rubens — ca 1625-1630)

The six language-related principles concerned with term- and phrase-processing before, during and after the translation have received the most focus from tool vendors and users in the past decade and a half. Perhaps surprisingly, it was terminology rather than segment-level translation that resulted in the first commercial products (TRADOS MultiTerm and TermStar from STAR). Although machine translation (MT) efforts date back to the 1950s, they initially did not have much to do with the language industry.

Infrastructure and term- and segment-level language processing clearly remain of basic importance today, but it is the last aspect — translation workflow and billing management — that is causing the most excitement and the greatest number of new products in the industry.

Let's first look at the six language-related principles in detail and see how we can match them with some of the past, existing or upcoming technologies and tools.

Term-level processing

Term-level before translation, the monolingual and bilingual term extraction for the creation of termbases and glossaries in preparation for translation projects, is probably the most overlooked area in practical terms. Many non-translation-related tools allow for indexing and concordancing of monolingual materials, but a surprising number of tools are also specifically geared toward the language industry.

Essentially, two kinds of technologies are used to achieve the extraction of terms, matching of term pairs and glossary creation: those that work primarily on a mathematical level ("if word *A* always appears in sentences for which word *B* always appears in the translated sentences, then these words must form a word pair") and others that work with underlying dictionaries and other language materials. Not surprisingly, the results from tools that use linguistic material are more accurate and superior to their competitors, but the number of languages that are supported is naturally more limited. Tools such as SDL PhraseFinder, TEMIS XTS, and Similis basically support English, French, German, Spanish, Dutch, Portuguese, and, in the case of XTS, a smattering of additional European languages.

It is anyone's guess why these tools are not being used more consistently, but I would assume that non-billable time plays a significant role in this.

Term-level during translation refers to the automatic terminology lookup that virtually all translation environment tools (TEnTs) perform. The term *TEnT* is used for the tools that are often referred to as CAT tools and that contain a translation memory (TM), a terminology maintenance, and a translation interface component. A terminology database, conceptually a highly customized dictionary that was either created in the previous principle or is created manually before or during the translation, presents project- or client-specific term pairs alongside other supporting information to the translation professional as he or she works through a text.

While all TEnTs offer this feature, the way they use it differs significantly. Not surprisingly, the two tool vendors that early on released their terminology components as standalone tools — STAR and TRADOS — have complex engines behind their terminology components, whereas many of the other tools use mere bilingual glossaries.

How the terminology component is integrated into the workflow also differs greatly and includes anything from a mere highlighting of source terms that are found in the termbase to the proactive display of matches to automatically entering them into (pre-) translated segments. Especially this last feature, coupled with the complexity or ease of entering the terms in the first place — unlike building up a TM, there is always a manual component of entering matches into a terminology database — makes a significant difference in how much the terminology component is used by the translator. The more immediate the benefit and the lower the cost (or the lower the entry or processing speed), the more likely it is to be used.

Term-level after translation refers to the terminology consistency check and the non-allowed terminology check. For this task there are both specialized tools (such as QA Distiller, Quintillian, ErrorSpy, or XBench) and increasingly many TEnTs that support these features internally. And in fact, it was the rise of these third-party tools that seems to have given the TEnT vendors the push they needed to implement quality assurance features such as white- and blacklisting terminology in the last two or three years.

From a technological point of view, a lot still needs to be done in this area. Some tools do use morphological rules for a limited number of languages aside from mere static matching rules, but most tools don't, and the majority of languages are not morphologically supported. So the translation professional more likely than not ends up being presented with long lists of mismatches that are due to the inability of the tools to recognize necessary morphological changes of target language terms or, for the same reason, non-flagged items on the source side.

Phrase-level processing

Segment-level before translation refers to the preparation that texts need to go through before TM or MT is performed. These include text segmentation, text alignment of previously translated documents and indexing.

Both text segmentation and alignment have recently seen significant improvements. Differences in text segmentation had long been the major obstacle to exchanging TM and corpus data between different tools. However, a relatively new standard, SRX — Segmentation Rules eXchange — has provided a major step toward overcoming this obstacle. Not all tool vendors have embraced this standard, and it is therefore not yet widely used, but within the next couple of years this should become something of a non-issue.

Alignment, long on the list of a translator's most undesired tasks, has also taken a quantum leap forward. Traditionally, alignment was performed by a pre-packaged feature of TEnTs that analyzed texts mechanically by segmentation rules and a limited number of non-linguistic markers such as numbers or styles.

However, through the recent commercial release of tools or services that specialize in alignment and use statistical (AlignFactory) or linguistic materials (AutoAligner), this task's accuracy and automation have made it once again reasonable to use alignment on large-scale projects. The resulting TMs can then be employed as traditional TMs or as training material for statistical MT engines.

Segment-level during translation refers to TM lookup and MT processes. TM lookup — the leveraging of content from translation memories and/or corpora — is at the heart of what TEnTs such as Across, Déjà Vu, Heartsome, JiveFUSION, Lingotek, MemoQ, MetaTexis, MultiTrans, OmegaT, SDLX, STAR Transit, Swordfish, Trados, Wordfast and any number of other tools do. In regard to this principle, a translator and a project manager may have different expectations: the translator is primarily interested in the ease and practicality of the translation environment; for the corporate user, workflow and translation management are of greater concern. These different expectations have put tool vendors in an awkward spot, not only resulting in different versions for the different user groups but also in shifting emphases on the different groups in different stages of the tool development cycles.

MT is experiencing its greatest revival since the excitement that surrounded it in the 1950s. And here again, different stakeholders have very different agendas. Though MT by and large remains the best-loved enemy of the freelance translator and is eyed with suspicion by smaller language service providers (LSPs), large LSPs and large translation buyers have long been running projects that are too large or too time-consuming for human translation through primarily statistical MT.

One of the more exciting developments today is partnerships between TM and MT vendors. This has the feel of the prodigal son written all over it, considering that TM was originally a subset of machine translation.

Segment-level after translation — missing segment detection and format and grammar checks — is the counterpart to the white- and blacklisting of terminology and terminology consistency checks on the term-level. And just as those developments were driven by the aforementioned third-party QA tools, so it is here. While few grammar checks are in place (aside from the tools that use the Microsoft Word interface), virtually all TEnTs provide a large variety of mechanical, non-language-related checks. These include missing segment detection and format checks, but also verification of numbers, punctuation and special characters. On a translation-related level, the introduction of QA features, including the term checks, has been one of the biggest pushes in the tool development during the last couple of years. And while it probably will take another year or two until they are accepted and widely used by the majority of translators, they are here to stay.

The workflow hurdle

Still, it's the last principle, translation workflow and billing management, that has thrown the language industry and its tool vendors into a tailspin lately. While there are certainly improvements to be made on the first seven principles, they are typically accepted as a given and are implemented in some way or the other on the translator's workplace or the LSP's network. But when sophisticated translation buyers who were used to complex software-based workflow and accounting applications took a more active role in the process, and when at the same time some global LSPs' growth could primarily be attributed to their sophisticated workflow products, there was an almost universal call for tools that would support these aspects of the business. Tool vendors responded with a number of different solutions.

Figure 1: St. Jerome has been forced to march to a new beat with different tools.

There are a number of powerful, web-based tools for LSPs such as Plunet, Worx and Beetext that cover various aspects of project management. Through partnerships or connectivity with TEnTs, these tools attempt to cover most technological aspects of the translation process.

Translation management systems (TMS) such as Idiom WorldServer and other corporate products from SDL and Across, along with any number of company-internal tools such as Lionbridge's Logoport and Elanex's ElanexINSIDE, also cover workflow and project management. However, these products essentially cover all eight principles: the infrastructure, the term- and segment-level processing before, during and after translation, and the translation workflow aspects. Though the market experienced some hiccups earlier this year when Idiom was acquired by SDL, this should be a market segment with significant growth potential.

The goal is in sight! Have we run through our translation technology survey in Olympic speed?

The cloud

Well, we're almost done. We have so far assumed that these principles refer to desktop-based or network-based computing, but they also apply to cloud-based computing. The internet has enabled translation technology users to collaborate and share resources — something that already seems natural in the days of Web 2.0 and beyond, but is still new in the lives of most translation professionals. Projects such as TDA, the TAUS Data Association, or TM Marketplace allow for the sharing of or access to data sources and at the same time open-source and commercial projects alike open to crowdsourcing. Tools such as Lingotek or the Google Translation Center offer translation interfaces that provide the necessary tools for translation and access to ever-growing public TMs. And these in turn can then be used to train MT to be more accurate.

To return to St. Jerome — he is without a doubt a fantastic model, but he has been forced to march to a new beat with different tools (Figure 1). If he continues to employ them wisely, they will not only make him more accurate and efficient, but give him access to a whole new world of communal resources and forms of collaboration. G