Data Standards: Can They Help Us and Can We Help Them?

By Alan Melby and Jost Zetzsche

Anyone who has ever taken the Trans-Siberian Railway between China and Moscow understands the importance of standards. Right at the border between China and Russia, the trains have to go through a "break-of-gauge" where carriages to or from China have to be lifted to have their "bogies" changed. (Bogies are structures underneath trains to which axles and wheels are attached.) The change of bogies certainly adds to the adventure of the long train ride; however, it is also a compelling illustration of the problems that arise from unaligned standards. Permit us to explain.

To address the problem of differences in railway gauges, the Parliament of the United Kingdom passed a law over 160 years ago (Gauge Act 1846) that defined the standard rail gauge as 1,435 millimeters.[1] This U.K. standard is used by over half the railway lines of the world, allowing rail travel nationally and internationally with a minimum of tedious bogie changes, but this standard is not universal. China uses 762 millimeters, New Zealand uses a gauge of 1,067 millimeters, and Finland uses 1,524 millimeters, the same as the former Soviet Union.[2]

Translators and Standards

How do railway standards relate to data standards in the translation industry? There are at least three ways. First, a similarity: standards are necessary to allow for smooth exchangeability between countries (or translation tools). Second, a difference: exchangeability within the language industry is not important enough for an authority like a government to step in and institute standards (and it is entirely possible that we would strongly object if one did!). Third, an observation: standards need not be completely universal before becoming highly useful.

Language industry veteran Kirti Vashee recently posted an entry on his machine translation blog about data standards in the translation industry. While some of the points in his blog entry need further discussion, we agree entirely with his vision of fully interoperable translation tools:

> I [should be able to] edit a document downstream with an application [translation tool] that did not create the original data and send it on to others who can continue the editing in other preferred applications [tools]. I think this is a big deal. I think this is the future, as data flows more freely in and out of organizations.[3]

Setting aside for the moment the problem of compatibility between the many different authoring environments in which documents are created, which is very important but much bigger than the translation industry, let's look at a translation-specific problem to illustrate the importance of standards in translation when using translation tools.

Are standards necessary? Yes, they make exchangeability a reality.

Text Alignment: Segmentation

The issue: how do you keep a source text and its "aligned" translation together from the starting point, where there is only a draft translation, to the end point, where the translation has been edited and proofed and is ready for publication? Just to be clear, we are talking about a *bi-text,* that is, a text and its translation that have been segmented and aligned so that each segment of source text is linked to the corresponding segment of target text.[4] Figure 1 provides an example of a very short bi-text.

Segments in a bi-text are typically sentences. True, they are sometimes paragraphs, but in this article we will focus on segments as sentences. We need not be aware of it, but most translation memory tools automatically create a bi-text while a translator produces a translation, working a segment at a time.

One way of making Vashee's vision into reality would be to have a standard format for representing a bi-text. Before a translation project begins, the project manager could make the source text into a bi-text file and send it to a translator. The translator would be able to choose from a variety of TEnTs (Translation Environment Tools), since they would all support the standard bi-text format.

There are many types of translations where segmentation is not as relevant and segment boundaries are often not retained. However, much translation work done by professional translators involves segment-oriented translation, with the occasional sentence being split into two sentences in the translation or two source sentences being combined in the translation.

As mentioned above, many translation tools already represent a text and its translation internally as a bi-text. The problem is that they also need to support a bi-text standard. In Vashee's vision, a translator can save a bi-text and pass it back to the project manager, who does not care which tool was used to produce it. Then the project manager can put the bi-text into some kind of quality assurance tool, such as one used for terminology consistency checking, and then pass it on to a bilingual editor (called a reviser in Europe), who can compare the source text and translation in a different tool than the one used by the translator and make changes as needed.

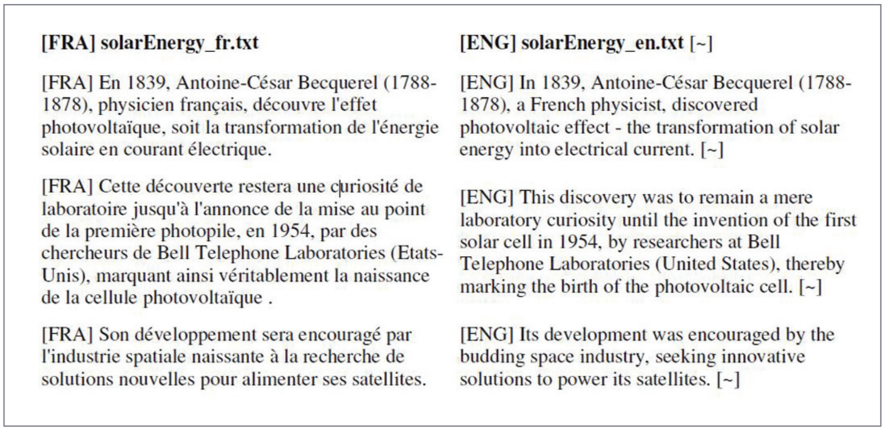
The source text and its translation would ideally remain in bi-text format all the way through to final proofing, although this would require support for the bi-text standard by word processing and desktop publishing software vendors.

The good news is that there is already a standard that can represent a bi-text. It is called XLIFF, and it is becoming more widely used. XLIFF is an XML-based standard that was originally developed for the bi-text representation of software files (XLIFF stands for XML Localisation Interchange File Format), but today it is used for virtually all file formats that can be processed by TEnTs.

While XLIFF represents the translation files, there is also the translation memory that contains translation units (pairs of translated segments) of previous and ongoing projects. The translator should be able to receive a relevant translation memory along with the source text and be able to read it in any tool and use it to identify segments that have been translated previously. Many translators will recognize that there is a standard that provides a degree of interoperability for translation memories: TMX (Translation Memory Exchange). Just like XLIFF, the TMX standard is also XML-based and is used to store, maintain, and exchange translation units between different TEnTs.

The reader may be wondering what

Figure 1: Sample bi-text



[FRA] solarEnergy_fr.txt	[ENG] solarEnergy_en.txt [~]
[FRA] En 1839, Antoine-César Becquerel (1788-1878), physicien français, découvre l'effet photovoltaïque, soit la transformation de l'énergie solaire en courant électrique.	[ENG] In 1839, Antoine-César Becquerel (1788-1878), a French physicist, discovered photovoltaic effect - the transformation of solar energy into electrical current. [~]
[FRA] Cette découverte restera une curiosité de laboratoire jusqu'à l'annonce de la mise au point de la première photopile, en 1954, par des chercheurs de Bell Telephone Laboratories (Etats-Unis), marquant ainsi véritablement la naissance de la cellule photovoltaïque .	[ENG] This discovery was to remain a mere laboratory curiosity until the invention of the first solar cell in 1954, by researchers at Bell Telephone Laboratories (United States), thereby marking the birth of the photovoltaic cell. [~]
[FRA] Son développement sera encouragé par l'industrie spatiale naissante à la recherche de solutions nouvelles pour alimenter ses satellites.	[ENG] Its development was encouraged by the budding space industry, seeking innovative solutions to power its satellites. [~]

XML is all about. Fortunately, when everything is running smoothly, a translator does not need to see any XML. It is used primarily for components of a computer system to talk with each other in computerese.

Together, XLIFF and TMX seem to be the basis for building Vashee's vision of the future. So why are we not yet there? Two obstacles: 1) more data standards are needed and 2) more tool vendors need to implement these standards. The need and a solution are discussed below.

Consistency in Segmentation

Just as with railway gauge standards, users do not notice departures from standards until there is a "break" of some kind. A huge break in data occurs if two tools need to be used together in a translation project involving translation memory but they do not both support XLIFF and TMX. However, a more subtle break can occur between XLIFF and TMX when segments are not defined consistently.

At first, it may seem that segmentation is a non issue. Isn't it obvious how to divide a text up into sentences? Generally, for a human, it is obvious. However, there are even cases where sentence boundaries are ambiguous for a human. Consider the following sentence (based on a sentence provided by Arle Lommel, the chair of OSCAR, the Localization Industry Standards Association's committee for the development of open standards).

> *Bill was forced to complete all the scraping, painting, finishing, etc. Bob was supposed to finish by Tuesday.*

Is this one sentence or two? This sentence could be paraphrased:

> *Bill was forced to complete all the scraping, finishing, etc., and so on, that Bob was supposed to finish by Tuesday.*

A two-sentence interpretation could be paraphrased:

> *Bill was forced to complete all the scraping, finishing, and so on. Bob was supposed to finish a different project by Tuesday.*

Such ambiguities are relatively rare, but there are many segmentation issues that occur frequently.

Translation technology developers Rodolfo Raya (Maxprograms) and David Pooley (SDL) were kind enough to share some of the segmentation issues they encounter in everyday work with XLIFF and TMX:

1. Should a semicolon be considered a signal of a boundary between two segments?

2. How about a colon? Sometimes a colon is followed by a list of nouns, but other times it is followed by another sentence. The difference is easy for a human to detect but not for a computer.

3. Does a tab indicate a new segment? At one point, two well-known translation tools differed on this question.

4. What should be done with "white space" (blanks, tabs, and new-line characters) that appear after periods? Should that white space be part of the segment or not? If one segmentation system retains the white space and the other deletes it, logically identical segments in a translation memory may not get "perfect match" scores in translation memory lookup.

5. The most obvious question has been saved for last: When is a period ***not*** the end of a sentence? To answer this question, the computer has to have a complete list of abbreviations. Of course, this list and other segmentation rules differ from language to language.

Suppose a text is segmented with one set of rules, translated, and put into a translation memory and exported as a TMX file. Further suppose that a slightly revised version of the same source text is segmented into an XLIFF file using a different tool and thus a different set of rules, and the TMX file from the earlier translation is then accessed. During translation memory lookup, some segments that remain unchanged that were previously translated will not be found in the translation memory due to segmentation differences. They will then have to be re-translated needlessly. In response to that scenario, an additional data standard, SRX (Segmentation Rules Exchange), was developed initially for use with TMX, but has been found to be applicable to all segmentation tasks.

An SRX file contains a formal set of rules for segmenting text. There may not exist one true set of segmentation rules for each language that everyone should use, but at least a translation memory (in TMX) can be accompanied by an SRX file that documents how it was segmented. Then, when a source text is segmented with the intent of leveraging TMX files against it, that source text can be segmented the same way the translation memory was segmented. However, this only works if all the tools involved in the project can export and import XLIFF, TMX, and SRX files.

Implementation of Segment-Related Standards

The Translation Tool Forum at last year's ATA Annual Conference included a document prepared by ATA's Translation and Computers (TAC) Committee with the cooperation of TEnT vendors that were exhibiting at the conference (Across, Atril, JiveFusion, Kilgray, Multiling, SDL, STAR, Terminotix, TotalRecall, and Wordfast). As of October 2009, all but one vendor had implemented TMX, a mere 4 out of 10 had fully implemented XLIFF, and only one had implemented SRX. Nevertheless, nearly all of these vendors indicated that they were planning to implement XLIFF and some were planning to implement SRX.

Clearly, the tool vendors are in a period of transition and this is where we, their customers, fit in. Would you like to encourage tool vendors to move ahead with an implementation of XLIFF and SRX? Then please send a message of support to datastandards@atanet.org. These messages will be compiled by the TAC Committee and given to the vendors.

The Future

We do not need to live with the equivalent of incompatible railroad gauges. Your influence on tool vendors—do not forget to send an e-mail supporting the implementation of XLIFF and SRX—can make the vision of interoperable tools a reality much sooner and avoid unnecessary loss of translation data.

Who Has the Data Standards for the Translation Industry?

While the U.S. government does not impose data standards on the translation industry, we do have a global association—**the Localization Industry Standards Association (LISA)**—that serves as an umbrella organization for many of the various committees that develop and publish the necessary standards. Another relevant organization is the **Organization for the Advancement of Structured Information Standard (OASIS)**, which has a committee on translation/localization.

OASIS Standard

www.oasis-open.org

XLIFF (XML Localisation Interchange File Format): The XML format for exchanging localization data.

Standards Originating from LISA

www.lisa.org

Segmentation Rules eXchange (SRX): The vendor-neutral standard for describing how translation and other language-processing tools segment text for processing. It allows translation memory and other linguistic tools to describe the language-specific processes by which text is broken into segments (usually sentences or paragraphs) for further processing.

Translation Memory exchange (TMX): The vendor-neutral open XML standard for the exchange of translation memory data created by computer-aided translation and localization tools. The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process.

Fortunately, there is coordination between LISA and OASIS on translation data standards.

References

1. Gauge Act of 1846. "An Act for Regulating the Gauge of Railways," www.railwaysarchive.co.uk/documents/HMG_Act_Reg1846.pdf.

2. *The CIA World Factbook, 2010 edition* (New York: Skyhorse Publishing, 2009).

3. Vashee, Kirti. "Are There Any Standards in Translation?" Retrieved August 18, 2010, from http://kv-emptypages.blogspot.com/2010/05/are-there-any-standards-in-translation.html.

4. Harris, Brian. "Bi-text, A New Concept in Translation Theory." *Language Monthly* (Volume 54, 1988), 8-10.