"Translating html the amateur way: double scrolling with highlighted markup"

Xavier de Pedro Puente. Technologies Area for Research. University of Barcelona. Pavelló Rosa. Travessera de les Corts, 131-159. Barcelona 08028 – SPAIN. xavier.depedro@ub.edu

Abstract

Amateur translators do not have the same experience as professional translators to move around the source and target documents without loosing too much focus and concentration. Thus, some aids are often very welcome to help those amateur translators in their everyday work localizing documents and keeping them in synchrony. This position paper describes a case where some freely available tools were used for some amateur translation job. These tools comprised simple yet powerful free software text editors with double or dual scrolling, spell checking and markup highlighting. Optionally, some Machine Translation (MT) resources were used in a first step for language pairs where high accuracy was reached by the system. Since the MT free software resources are getting better with time, such as the case with Apertium or Bitext2tmx, the post-editing in double or dual scrolling software would be more helpful year after year, even for language pairs not so related.

Context

It seems to be commonly understood that professional translators do not need computer programs with dual scrolling of source and translated documents. In fact, they could find it even annoying (Desilets, personal communication). This could be due to the fact that they are very good at quickly orienting themselves around text, in the same way that developers are very good at orienting themselves around code. However, professionals from other knowledge areas that need to do some translation for a multilingual audience might eventually work as amateur translators and they might not be that experienced moving between versions of the same document in different languages. Therefore, some aids to ease those tasks are often very welcome. Moreover, in a collaborative translation context, translators are often amateurs, as opposed to professionals in other contexts, and they may need different types of tools (in particular free ones [0]). Thus, this position paper describes the solution adopted in one of those scenarios using freely available solutions, mostly based on free software which can be run on most operating systems.

Main handicaps for (at least) amateur translators

In cases where the language knowledge is not the biggest handicap for the amateur translator, the major issue to be solved usually is finding the place where the translator was in the source document, each time he wants to come back after the next bit is translated in the target document. This iterative process of looking back and forth the source and target document in process of translation usually produces eye fatigue and concentration lost, as far as it frequently comprises moving the eyes, head and body position a lot at each iteration until finding the right place in every step.

Most amateur translators tend to use a paper copy of the source document, and a computer version of the target document which contains the translation. But this means printing big amount of pages for translation, when the source documentation is long, and changing the page face from time to time, while loosing focus and concentration again on the translation process.

Moreover, sometimes the translation process seems to be accelerated if some machine translation (MT) resources are used prior to the human translation effort, whereas in some other cases, the amount of time required for "post-editing" (i.e. to refine the previous automatic translation; [1]) might be higher than what you would have spent translating the source document directly from scratch into the target language. This is an important issue where more scientific research is needed, as the MT resources improve year after year, even for not so similar languages such as Spanish and English [2][3].

Translating through source html with markup highlighted:

One specific scenario that the author of this position paper had to deal with was producing a new version for a technical manual of a new version of "*Curricul@*", a corporate software program to manage researchers' Resumes [4]. This new version of the manual had to be in three languages: in Catalan and Spanish, updating previous documents, and a brand new document in English. Source document was in html and divided in small sections which were translated one by one, sometimes updating the outdated version of the translated text, and sometimes from scratch. In this case, a tool like the Cross Lingual Wiki Engine (CLWE, [5]) would have been very helpful, but it was not yet ported to that corporate software by the time of this writing.

When help was needed from online MT resources (Apertium, Google, ... Figure 1), html tags were previously replaced by similar tags with spaces in the middle, so that they were not parsed but kept along the MT and easily converted back to html tag with simple search and replace rules post MT.

Figure 1: Translating short html document online with Apertium. http://www.apertium.org/

The post-edition was performed using a simple (yet powerful), multi-platform text editor called "Kate", from the KDE desktop software (Figure 2), on GNU/Linux. This free software small

program allows dividing vertically or horizontally the window in two columns or rows, corresponding to two different texts which can be scrolled individually, as well as it allows spell checking and markup highlighting.

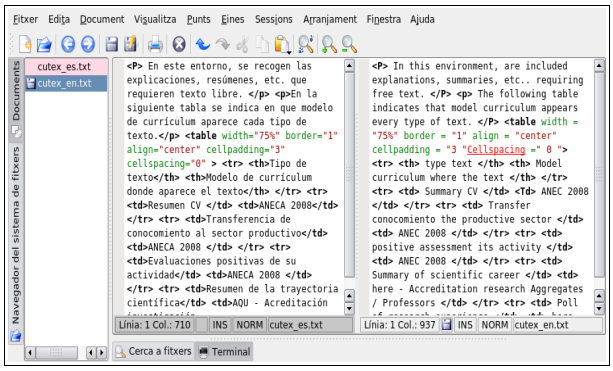

Figure 2: Double scrolling source and target documents on a simple multi-platform text editor (Kate, from KDE)

“Post-edit” translation

Our experience in this case indicates that, when the two languages are very similar and, therefore, it is easier to take advantage of a good translation efficiency, the MT plus post-edition [1] seems to be worth the effort, since it correctly translates most of the document among both languages. This is the case of Apertium with Catalan – Spanish or other language pairs, with error rates just between 5% and 10% [2].

On very different language pairs (Catalan <-> English, Spanish <-> English, ...), much lower efficiency on MT is expected, and thus, post-editing work would be too much compared with translating from scratch.

Double scrolling vs. dual scrolling

The double scrolling allowed by free software programs like Kate produces the same effect as when you work with dual monitor setup in your computer, keeping them both attached one next to the other. When you move along the translation, you have to “double scroll”, i.e., scroll on both screens individually.

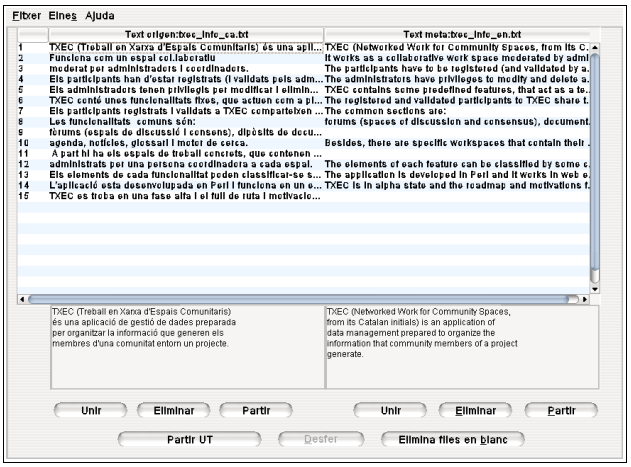

Figure 3: Bitext2tmx showing a text in two languages with a single horizontal and vertical scroll

There is another approach for such task which is starting to be found on free software multi-platform solutions: “dual scrolling”. It comprises one scroll bar which controls both screens at the same time, in a similar way as some “diff and merge” free software programs work (kdiff3, winmerge, ...). One of such programs is “bitext2tmx” [6], java based, which in addition to dual scrolling (Figure 3), it includes the chance to generate a translation memory, in TMX format, from both text documents for use in computer-assisted translation applications. However, the author must admit that even if dual scrolling seems to be a promising tool to help translation process, the current early stage of the software did not allow to make the translation easier than using free text editors with double scrolling, as previously reported. The main issues with “bitext2tmx v1.0” were four: (1) Syntax highlighting is missing, (2) as well as spell checking, (3) you need to move your eyes up and down too frequently to switch between selecting lines from the list (above) and reading the full line texts on both languages (below), and (4) you cannot use just the keyboard for the translation process, since the mouse is needed to switch the prompt between the upper and lower section of the window. Instead, enabling a behavior like the one allowed by the program shown in Figure 2, provided that dual scrolling was also possible, could be more helpful and less eye tiring, from an amateur translator point of view.

Conclusion

Even if some better designed research is needed, it could be the case that machine translation post-edition is worth the effort only for for language pairs where high accuracy was reached by the system. Another factor that contributes to make it worth is the presence of some markup such as html in the source document which allows the user to recognize more easily the highlighted or colored landmarks of the translation process back and forth between the source and target document. Dual scrolling in its current state was found to be less useful than expected, and double scrolling software was primarily used instead.

Since the MT free software resources are getting better with time, such as the case with Apertium or Bitext2tmx, the post-editing in the double or dual scrolling software program would be more helpful year after year, even for language pairs not so related.

Bibliography

[0] Free Software (Wikipedia entry). . http://en.wikipedia.org/wiki/Free_software. (Visit: 30/05/08)

[1] Allen J. 2003. **Post-editing**. In *Computers and translation: a translator's guide*. Somers H (Ed.). pp. 297-317.

[2] Armentano-Oller C, Corbí-Bellot AM, Forcada ML, Ginestí-Rosell M, Bonev B, Ortiz-Rojas S, Pérez-Ortiz JA, Ramírez-Sánchez G & Sánchez-Martínez F. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *Osmatran: open-source machine translation, a workshop at machine translation summit x, phuket, Thailand*. [Online: http://dlsi.ua.es/~mlf/docum/armentano05p.pdf]

[3] Armentano-Oller C, Corbí-Bellot AM, Forcada ML, Ginestí-Rosell M, Montava Belda MA, Ortiz-Rojas S, Pérez-Ortiz JA, Ramírez-Sánchez G & Sánchez-Martínez F. 2007. Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática. In *Proceedings of the floss international conference 2007* [Online: http://dlsi.ua.es/~japerez/pub/pdf/flossic2007.pdf]

[4] GREC Applications, 2008. Curricul@ Manual. 37 pp. University of Barcelona. http://gclub.ub.es/dl77

[5] Huberdeau L, Paquet S & Désilets A. 2008. The Cross-Lingual Wiki Engine: Enabling Collaboration Across Language Barriers. In *The International Symposium on Wikis (WikiSym)*. Porto, Portugal.

[6] bitext2tmx: Cross-Platform Bitext Aligner. Forcada, ML; Martin, R. http://bitext2tmx.sourceforge.net/ (Visit: 13/07/08)