

Research on translation tools

IGNACIO GARCÍA

University of Western Sydney, Australia

This paper responds to two of the three questions presented at the Tarragona symposium in relation to research on translation tools: 1) What specific problems need to be addressed by research?, and 2) What specific methodologies are required?

What specific problems need to be addressed by research?

In this paper, the term “translation tools” means computer applications designed specifically to help with the translation of digital content. It includes applications to help translators (mainly translation memories, or TM), applications to help corporations and institutions have their content translated (translation management systems, content management systems), and applications to help computer users understand unfamiliar foreign content (machine translation, MT). While our present focus is on research issues in relation to TM, when required we will also deal with how translation management systems and MT impact on TM.

Issues already tackled by research

Before delving into what kind of research should be done in the area of translation tools, let us consider what has already been done or is currently in progress. For convenience we will divide empirical research on TM into descriptive and (quasi) experimental studies. The first group includes surveys on TM adoption, the best known being those conducted by the Localization Industry Standards Association (LISA 2002, 2004), by eCoLoRe (2003) and by Lagoudaki (2006). We may also include reports, case studies and other research written by industry bodies such as LISA, Byte Level Research, Common Sense Advisory, and Translation Automation User Society (TAUS).

For the (quasi) experimental group, we note just a few examples of the issues addressed by researchers up to now. Work has been done on segmentation, with Dragsted (2005) finding that expert translators use longer translation units than novice translators do, and that the segmentation rules applied by TM editors do not correlate well with the way expert human translators “chunk” a source text (2004, 2006). There has also been some investigation of the cognitive effort required to deal with TM fuzzy matches

and with MT post-editing (O'Brien, 2006a, 2006b). Guerberof (2008) has also compared fuzzy matches with post-editing, but with a focus on time and quality. TM output quality has been investigated by Bowker (2005) and Ribas López (2007), who both seeded translation memories with errors to check whether translators would detect them or, conversely, the memories would act as vehicles for error propagation. There have been attempts to assess the different ways of translators accessing memories, with Wallis (2006) finding that although productivity was similar translators seemed to prefer working in interactive translation mode (finding matches from the memory one segment at a time), rather than in the pre-translation mode (on a bilingual file in which all exact and fuzzy matches had been inserted).

New areas for research in a rapidly changing environment

One risk for those seeking to perform TM research is that, with the pace of technological change being so rapid, by the time the research is published the issue in question may no longer be relevant. To some extent, this is what happened with Willis (2006): the pre-translation mode, then prevalent, is now being phased out in favor of the emerging web-interactive mode in which translators find matches one segment at a time, but from remote databases (García 2007).

The movement of databases from the desktop to the server is not the only major development that has occurred in the translation industry over the past two or three years. There are another three important phenomena of interest to us here, namely the availability of massive databases, the convergence of TM and MT, and the larger role of MT. There are also two new pressures impacting on the industry and significantly they are coming from outside the industry rather than from within: we have termed them “hive” translation, and translation as “utility”. Each of these technologically driven developments opens brand new areas for research.

From the hard-drive to the server

Memory and terminology databases are being moved from the hard-drive to the server, and this is not just something only the localization giants (Lionbridge, SDL) can afford. Any medium or even small language vendor (or language buyer) can configure databases for remote access, and even budget TM tools (e.g. Wordfast) now permit it. This is changing the working conditions of translators, and anecdotal evidence already suggests many are not happy about it. Foremost, translators lose control over the resources they generate, while another common complaint is the slow turnaround time in opening and closing segments remotely via the server.

As just one example of these new issues that fresh research should address, Wallis's work (2006) could be updated by comparing productivity and user-friendliness of pre-translation mode against the web-interactive mode.

Availability of massive databases

Some TM tools already come with built-in databases (Lingotek) or allow direct access to them (viz. Wordfast's Very Large Translation Memory, or VLTM). How useful these are for direct segment matching remains an open question, but they are certainly useful for "concordancing"—i.e. searching for translation context at the sub-segment (terms, phrases) level.

It is not only software vendors who are interested in massive databases. The idea of compiling and sharing them has also reached corporations and institutions. Traditionally wary of freelance translators using their glossaries and memories when working for other clients in the same industry, corporations now seem to be realizing that they have more to gain by pooling linguistic resources together. This is what TM Marketplace pioneered from 2005 onward, and what the TAUS Data Association (TDA) is attempting now. These massive databases would then feed both TM and MT engines.

We can identify two approaches to accreting these massive databases. The Lingotek and VLTM model relies on crawling the web for bilingual text to align, without much attention to the quality of the material (often the search will include pages that have been poorly translated, or machine translated). On the other hand, the TM Marketplace and TDA model uses databases that include only copy and translation that has been published or vetted by known reputable sources. Interestingly, both methods rely implicitly on some form of vetting by expert human translators. In the former, loose segments will be ranked over time as translators choose them for re-use, while in the latter the segments are pre-vetted before release, whereupon they will be presumably subject to further vetting/ranking as they are accessed.

This immediately suggests one useful area for research: given the inevitable presence of post-vetting, does the extra effort required in compiling carefully pre-vetted bilingual databases pay off, or is the quick and dirty shotgun approach ultimately as productive? Further, can we generalize, or are the different approaches suited for different types of texts and/or tasks, and if so, what are they?

Convergence of TM and MT

A few years back it could categorically be said that MT was language-specific while TM was not; that MT came with sets of language specific-rules and vocabularies while TM came as a kind of empty receptacle into which translators poured sentences and terms. Nowadays, of course, some

TM tools come preloaded with data, as is the case of Lingotek and Wordfast. Others come with language-specific rules that promise greater productivity by facilitating re-use at sub-sentential level (Similis, Masterin). Furthermore, TM output is directly enabling the surge of Statistical MT, while the use of a hybridized TM/MT approach (so-called Machine Translation Memory, or MTM in the Idiom Word Server parlance) is now becoming standard practice in the localization workflow. We thus have another possible topic for research: Do these new language-specific tools work better than the tried-and-tested traditional ones (Trados, Déjà Vu, etc.), and who benefits most (translators, localizers, clients)? As for the integration of TM with MT or, better, the ingestion of TM by MT, that deserves a section apart.

Integration of MT with TM

MT, whether rule-based (RBMT), statistical (SMT) or hybrid, has gained a much higher profile over the last few years. Google Translate allows the user to set their browsers so that MT of any web page is just a click away. Microsoft Live Translation is configured so that when users demand a previously untranslated article from the Knowledge Base, they will receive an MT processed version. However, it is not this type of user-driven unassisted MT that concerns us here, but rather MT integrated with TM for use by translators, and with controlled authoring, TM and human post-editing workflow for the enterprise.

In the past, several TM tools came with plug-ins for MT, but the idea did not gain acceptance at the time. The prevailing view was that MT input would only distract translators, who preferred working on blank TM target-text segments. However, SDL Trados 2007 again offers this possibility of filling in blank segments with MT hints. Given the advances in MT since then, can this feature actually be useful this time around?

A recent survey conducted by SDL (2008) found that almost one quarter of the clients surveyed were already using MT for some out-bound translation tasks, or were considering using it. For localization tasks, systematic use of MT is not an if, but a when. The role of translators is steadily moving away from checking TM matches and completing blank segments, into straight-forward MT post-editing, and this is one of the most exciting areas of translation technologies research right now. The key question for localization now is when, under what conditions, and for what type of task, controlled language plus TM plus MT plus post-editing will produce equal quality faster and more cheaply than the current TM model. O'Brien (2006a, 2006b) and Guerberof (2008) have already commenced investigation into this area, but much more work is needed because of the profound impact this change will have on translators, translation commissioners, users and trainers.

Other technology-driven change in translation

Technology is clearly shaping the translation industry (and the way individual translators work) from within by making more efficient tools and processes. However, perhaps with more far-reaching consequences, technology is also impacting on the industry and on translators from the outside, via two emerging trends that we have termed “hive” translation and translation as a “utility”.

By “hive” translation we mean the outsourcing of web content translation to bilinguals within the community (TAUS calls this “community translation”, but the unbounded nature of cyberspace associations transcend old notions of “community”). While such “crowd sourcing” is not new—it is the cornerstone of translation within the free and open source software (FOSS) sector—we now find it being applied to commercial work that under normal circumstances would be dealt with by paid professionals. This new trend is best exemplified by the well-known social-networking site Facebook, which is being localized by its own users.

Then there is translation as a “utility”, which we could also perhaps call “translation-on-tap”, or “off-the-wall”, by analogy with public utilities such as water or electricity. This is best illustrated at present by Livetranslation.com, which offers fast, small-volume, user-friendly human translation-on-demand. Here the client posts a source text to the site and, with arranged payment, a duty translator performs the translation and uploads it in the time it takes to type it. This could be ideal for email and social networking content for which unassisted MT has not produced adequate output. It may sound trivial, but Microsoft is taking it seriously enough, with plans to configure its Knowledge Base so that if the users are unhappy with the results from its unassisted MT engine, they can access this premium “human” service.

Although both these initiatives are still in the trial phase, they deserve close attention because, whether taken separately or in combination, they have the potential to dramatically change the way translation will be performed (and consequently assessed and researched) over the coming decade.

What specific methodologies are needed?

The studies mentioned in 1.1 above offer methodological approaches that could be validly applied to the fresh fields being opened by technological advance. On the descriptive side, more surveys are needed to map the changing trends of TM/MT uptake. With industry cooperation, case studies could show for example how content previously translated using TM has been shifted into new work flows involving MT plus post-editing, and with what results.

Ethnographic methods, including self-ethnography, have been scarcely utilized. By allowing researchers to explore processes that they themselves may have been instrumental in implementing, we could achieve some interesting findings. A professional background in localization, for example, could be a great advantage for performing research in this area.

The experimental method could be adapted to correlating TM and MT outputs, process/product patterns of professional and amateur translators, the usability traits of quick and dirty massive databases versus well-vetted ones, etc. Indeed, the range of variables for study could be as broad as the ingenuity of the researcher.

The aforementioned and similar studies on the larger area of translation and revision illustrate some of the tested ways of gathering data for analysis. The means can be linguistic, such as think-aloud protocols (Krings 2001, Kunzli 2006), or employ technology-rich instruments such as keyboard-logging (Dragsted 2005, 2006, Jakobsen 2002, O'Brien 2006a, 2006b) and eye-tracking software (O'Brien 2006b). The resulting data could be successfully applied not only to cognitive processes (such as correlating pauses with translation-unit boundary markers, or with signs of cognitive load), but also to usability studies concerning which tools (or approaches) are more productive or user-friendly on the basis of translator interaction with the applications and end-user interaction with the translated text.

The forthcoming availability of massive databases should allow for corpus-based research to cross-contrast the text patterns found in human, machine-aided and automated translation, and perhaps allow us to start delving into the critical question of which text types would be best suited to which translation mode.

As we move into the second decade of the millennium, modern computers are making it easier to handle the research instruments needed to probe texts and to inquire into how translators and end-users process them. Since we are visibly reaching a tipping point in the use of technology for translation, the research we undertake on translation tools now will not only be relevant and exciting for us, but seminal for future users and investigators.

References

- Bowker, Lynne. 2005. "Productivity vs quality? A pilot study on the impact of translation memory systems". *Localisation Focus* 4(1), 13-20.
- Dragsted, Barbara. 2005. "Segmentation in translation: Differences across levels of expertise and difficulty". *Target* 17(1), 49-70.
- Dragsted, Barbara.. 2006. "Computer-aided Translation as a Distributed Cognitive Task". *Pragmatics & Cognition* 14(2), 443-464.

- eColore. 2003. "Translation Memory Survey"
<http://ecolore.leeds.ac.uk/xml/project/news.xml?lang=en>. Visited November 2008.
- García, Ignacio. 2007. "Power shifts in web-based translation memory".
Machine Translation 21(1), 55-68.
- Guerberof, Ana. 2008. "Post-editing MT and TM. A Spanish case".
Multilingual 19(6), 45-50.
- Jakobsen, Arnt Lykke. 2002. "Translation drafting by professional translators and by translation students". *Empirical Translation Studies: Process and Product* 27, 191-204.
- Krings, Hans. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Trans. G. Koby, G. Shreve, K. Mischericow & S. Litzar. Ohio: Kent State University Press.
- Kunzli, Alexander. 2006. "Translation revision - A study of the performance of ten professional translators revising a technical text". In M. Gotti and S. Šarčević (eds.) *Insights into specialised translation*, MBern/Frankfurt: Peter Lang. 195-214.
- Lagoudaki, Elina. 2006. "Translation Memories Survey 2006". London: Imperial College London. http://www.atril-france.com/index_fichiers/Docs/Comparaison_logiciels_TAO.pdf. Visited November 2008.
- LISA. 2002. "Translation Memory Survey".
<http://www.lisa.org/products/surveys/tm02survey.html>. Visited November 2008.
- LISA. 2004. "Translation Memory Survey".
<http://www.lisa.org/products/surveys/tm04survey.html>. Visited November 2008.
- O'Brien, Sharon. 2006a. "Controlled Language and Post-editing".
Multilingual - Writing for Translation - Getting Started Guide 17(7), 17-19.
- O'Brien, Sharon. 2006b. "Eye-Tracking and Translation Memory Matches".
Perspectives: Studies in Translatology 14(3), 185-204.
- Ribas López, Carlota. 2007. "Translation Memories as Vehicles for Error Propagation. A Pilot Study". Unpublished Minor dissertation, Rovira i Virgili, Tarragona.
- SDL Research. 2008. "Trends in automated translation in today's global business. White Paper". <http://www.sdl.com/en/globalization-knowledge-centre/whitepapers/>. Visited November 2008.
- Wallis, Julian. 2006. "Interactive Translation vs. Pre-translation in the Context of Translation Memory Systems: Investigating the effects of translation method on productivity, quality and translator satisfaction". Unpublished MA in Translation Studies, University of Ottawa, Ottawa.