Walter Daelemans and Véronique Hoste, eds. *Evaluation of Translation Technology*. Special issue of *Linguistica Antverpiensia New Series — Themes in Translation Studies* 8. 2009. 261 pp. ISBN 9789054876823; ISSN 0304–2292. 35 euros.

Reviewed by Anthony Pym (Tarragona)

As is increasingly common these days, this is at once a collective book and a special issue of an academic journal. It is on a specific and important topic; it has some excellent empirical research papers by some of the leading scholars in the field; yet I am not sure we are well served by the hybrid format of a journal that is also a book. To understand why, you have to know what is there.

The volume opens with a blast: Andy Way argues that statistical machine translation, which would be pointing the way forward, was pioneered by people so convinced of their vision that they never bothered to make their work understandable to other Machine Translation scholars, let alone to the rest of us. The argument hinges on colorful descriptions of what happened at key conferences, who said what, who remembers what was said, where the main players were seated at the time, and presumably how the seating affected what was remembered. The main proposition was, and remains, that if the aligned databases are big enough, the statistical probability of co-occurrence will select better renditions than will the writing of linguistic rules for each language pair. This is potentially revolutionary for two reasons: the statistics should work for any language pair (if the databases are there) and the more the system is used, the bigger the databases, the more exact the statistical matching, and the better the translations. That is, the more you use statistical machine translation, the better the translations become. And that could have been the end of the day for linguists, long ago. As IBM's Fred Jelinek is reported as saying in the years around 1988, "Every time I fire a linguist, my system's performance improves" (23). Way points out that Jelinek was actually talking about speech recognition, but the terms of the debate, and the values at stake, are clear enough.

Way's main historical point is that the statistical revolution in machine translation is now not as radical as it was made out to be, and that hybrid systems (incorporating syntactic modeling) are giving good results. The pleasure of reading the article, however, lies in Way's *aperçu* into the people behind the machines, with all their individual brilliance and petty rivalries. The technical field is thus humanized. Way then helpfully attempts to explain the technicalities at stake, with a result

Target 25:2 (2013), **286–290**. DOI 10.1075/target.25.2.08pym
ISSN 0924–1884 / E-ISSN 1569–9986

that is rather less understandable (the absence of key symbols on p. 26 doesn't help much). In all, this is an excellent keyhole look at the recent developments of machine translation, for specialists and non-specialists alike. As Way presents it, the key issue being played out is the relative advantages of statistics-based machine translation. The strange thing is that, for the rest of this book, that particularly and powerful central debate is scarcely on the horizon, and the claims of the personalities are reduced to a world of merely technical terms.

Way is followed, as it happens, by a straight descriptive article on the web-based Framework of the Evaluation of Machine Translation developed within the context of International Standards for Language Engineering. The text reads like an instruction manual for someone using the framework to select the system they want, which is ultimately not much more complicated than putting in all the criteria for buying the exact car that you want (model, color, accessories, price, etc.). Does anyone really buy a car that way? Does anyone really invest in machine translation that way? Surely there is a whole world of symbolic values, bets on futures, and words-of-mouth that influence such decisions, not to mention the colorful personalities and outright bad-mouthing mentioned by Way in the previous article? As Iulia Mihalache argues in a later article in this same book, the use of one technology or another significantly involves the social process of entering one community or another, with the multiple subjective factors thus entailed. We are not all as straightforward as the engineers' instruction manuals would have us be.

Much the same could be said, I suspect, for the following article, where Vincent Vandeghinste takes us through the fairly dry and technical paces of actually setting up a hybrid machine-translation system. This should surely be published in a more specialized journal, where the language engineers are likely to find it and perhaps delight in the details? At this point, you see, I was starting to wonder exactly who the reader of this volume should be. And what happened to the apparent revolution in machine translation?

There then follow three genuinely excellent papers, all reporting on empirical research projects, all accessible to a wide range of readers (linguists and technicians), all hopefully useful in one way or another, and all presenting new information that deserves to be read and cited near and far.

Bogdan Babych and Anthony Hartley set out to give automatic indications of whether multiword expressions (e.g. "washing machine", "meet the demand") are systematically mistranslated by machine-translation systems. This is first done with rule-based Systran 6.0 between English and French/German/Spanish. The BLEU (mechanically calculated) evaluation metrics for 260 multiword expressions are then ranked in accordance with the risks they entail, where risk is measured in terms of both the frequency with which the expressions occur (thus the occasions for possible error) and the high potential impact of any mistranslation

(measured in terms of a low BLEU score). Linguists who write rules for machine-translation systems can then focus first on the expressions presenting the highest risk, presumably starting with the ones that are both frequent and of high impact (although pp. 94–95 seem not to elucidate how the two variables are weighted to give the priority index). There is a certain practical beauty in this: one does indeed see how this mode of calculation can be used to speed up the writing of rules. But what if statistics have done away with the need for rules (as was the basic proposition back in Way's world of colorful characters)? What if the linguistics is not really necessary? Babych and Hartley then go further and compare the Systran performance not just with human expert translation (which does better, of course) but with four other machine translation systems, including one that is statistically based. That is the point where I would like to know which system does best, and why. The authors, however, seem to shy away from the comparisons: the numbers say that Reverso did better than Systran, and that the statistics-based Candide was actually pretty crumby on this occasion. For all that, the authors only want to focus on Systran, and we are left wondering why. Is it simply because Systran incorporates linguistic rules and thus keeps the linguists in jobs? So much for the major debate of our time.

Nora Aranberri-Monasterio and Sharon O'Brien similarly use Systran (here 5–05) to look at how *–ing* forms (e.g. "viewing") fare when rendered from English into German, Spanish, French, and Japanese. They find Systran actually does pretty well, getting 72–73% correct renditions for all languages except French, which was at around 50% because of particular problems with two structures. The authors then compare human evaluation with five different metrics that use short constituents (thus excluding the BLEU metric). The evaluations by all these metrics were found to correlate strongly with the human evaluation, for all target languages, which is good news. Perhaps the best news, however, is that the analysis locates particular *–ing* forms that are more problematic than others. Linguists can thus focus on those particular cases (unfortunately not characterized by risk), either by writing rules for their translation or, probably more profitably, transforming the structures by editing source texts ("post-edit the source text"). So the linguistics is still there, in the absence of any testing of statistics.

All the above authors share a standard equivalence-type view of translation quality, described by Aranberri-Monasterio and O'Brien in terms of "a grammatical text which transferred the same information as the source text" (111). Lynne Bowker, in her article, effectively questions the linguists' presuppositions of what a good translation is. She reports on surveys of users' evaluation of machine translation in Canada: one survey looking at French to English (for the minority of English-speakers in western Quebec), the other at English to French (for the minority of French-speakers in Saskatchewan). The interest, says Bowker, is

that the demand for translations is greater than the supply, thanks in part to a "recognized shortage of professional translators in Canada" (130). What is really interesting, however, is that when the two communities of users were asked to assess the translations "keeping in mind the associated production time and cost" (142), they manifested very different opinions of machine translation. The French users, for whom translation was more a matter of protecting their language and culture rather than making information accessible, had generally low opinions, whereas the English-speaking community in Quebec, which seems more reliant on translation for access to information, gave more positive evaluations. This difference is not just due to the different minority situations of the two communities: 48% of the French-speaking respondents said they were language professionals, as opposed to just 9.2% of the English speakers. That is, perhaps, non-linguists are more tolerant of cheap-and-dirty translations than are linguists. Or they are more concerned about the social labor required to produce high-quality hand-made translations. This should ideally reflect back on the equivalence-type quality concept used in the previous articles, where linguists might be simply protecting their turf. Or again, it might explain why Bowker, like the other authors here, works with rule-based machine-translation systems (her experiment was prepared with Reverso Pro), not only avoiding the question of statistics but also removing from her evaluation the prospect that a system can be better thanks to constant intelligent use. Instead of being passive evaluators, the users might also be participants in a system's development. But that is not the debate here.

The remaining articles are not bad but could perhaps be elsewhere, since they have little in common and only one really addresses the question of evaluation. Iulia Mihalache runs through a series of "sociological, economic, organizational, cultural and psychological perspectives" (p. 176) on the way the use of technologies involves joining and transforming communities, all illustrated by citations from various web discussions. The purely qualitative evidence cannot be wrong, which means that none of the models is shown to be wholly right. Alberto Fernández Costales points out how the program Passolo can help you localize software, but since there is no comparison with anything else, we are back to the discourse of instruction manuals. Lieve Macken is more properly empirical in her comparison of sentential and sub-sentential segmentation, and uncontroversial in finding that a sub-sentential system gives better performance for text types with shorter sentence and more standard terminology. This is comforting but unrevolutionary. And a final article reports on a corpus-based comparison of 40,000 original and localized webpages in Spanish — casi nada —, only to fall into all the traps of using corpus analysis to make suppositions about process. Since it is assumed that "all Web texts are localized using TM tools" (213), the greater variation found in the localized pages is attributed to the use of those tools. Unfortunately, several

hundred processes can enter into the localization of websites, and the greater variation discovered here could also be a result of any number of them, or the act of translation itself, or more probably the great range of target audiences and thus varieties of Spanish being used. In short, there are too many unknown factors for any causal connections to be made, and the inclusion of this article is ultimately embarrassing. Needless to say, none of these articles address the ostensibly key debate over the evolution of statistical machine translation.

I think a good academic journal, in the traditional sense, might have included Way and the three strong empirical articles, the ones that report on new empirical research. The others would then be in different places, for different readers (instruction manuals, or more specialized journals). As it stands, however, the attempt to produce a book brings together texts that seek quite different readerships, with remarkably few crossovers, and connects poorly with the 20 pages of unrelated book reviews that then follow. Indeed, the problems of specialization might be evidenced in the editing problems involving mathematical symbols (pp. 26 and 83), suggesting that the copyeditors were not quite following what was happening.

A huge amount of work is being put into making *Linguistica Antverpiensia* a first-class international Translation Studies journal. All that effort, however, can run into trouble when a journal becomes a series of books.

Reviewer's address

Anthony Pym
Intercultural Studies Group
Universitat Rovira i Virgili
Ave. Catalunya 35
43002 Tarragona
Spain

anthony.pym@urv.cat