Continental Shifts?

Pangeanic is a mid-sized Spanish language service provider that started relatively early (2008) to use machine translation for its own production processes. Like most other statistical machine translation systems, it was based on the open-source *Moses* system. (By the way, have you ever wondered whether the makers of *Moses* realized the irony that their product's namesake had a speech impediment?) Pangeanic kept on refining the product with a more user-friendly interface and other options and started to market it in 2010 to other companies (under the product name *PangeaMT*).

About a third of their customers are other language service providers, and the other two-thirds are corporate customers. That's something Pangeanic would like to change -- by specifically focusing on a new segment of the market: freelancers. Freelancers are clearly the group in the translation industry that has been least targeted by MT vendors (and with some freelancers, that's quite all right!).

There are rules-based machine translation systems like *Systran* and *PROMT* (both of which actually have a statistical machine translation component as well, but only in the high-priced server editions that are essentially not available for freelancers to own); rather involved packages like *Do Moses Yourself* by Precision Translation Tools; companies like *KantanMT* that offer statistical machine translation on top of already pretrained machine translation engines (but start at 500 euro per month); SDL's *BeGlobal* at a minimum cost of $1,500 per year (for three million translated words); and lastly *Microsoft Translator Hub,* which is free in exchange for donating your (or make that: your client's) data. (Premium subscribers can read about all these different tools in the indexed Tool Box newsletter archives.)

So, pretty slim pickings for a freelance translator. The solution that *PangeanMT* now offers to freelance translators is -- as far as I can tell -- strictly paid on a per-word basis (.0002 euro cents per word) with no cost for training or retraining the cloud-based machine translation engines. Since the training of the machine translation engines in the available language combinations (EN<>ES, FR, IT, DE, DU, PT, NO, SV, and JA, plus Chinese starting in September and Arabic and Russian at the end of the year) is done on top of already-trained machine translation engines (based on TAUS and EU data), there is no minimum amount of data that needs to be added to achieve some quality improvement -- but clearly you won't achieve much with a couple hundred translation units. (For reference purposes: the makers of *KantanMT* recommend a minimum of one million words -- I wouldn't be surprised if that's about the same here.)

There are several tools that *PangeanMT* offers for cleaning the TM data before you use it to train the machine translation engine. This is an interactive process where you are prompted to look at potentially offending segments and either delete them or use them to improve your glossaries (which are also a component that is used in the training process).

Manuel Herranz, whom I talked to about this, was refreshingly realistic about some of the limitations of the system. For instance, he was very cognizant of the fact that not all language combinations always produce the same kind of quality. German lags behind the Scandinavian and Romance languages, Russian is even further behind, and Japanese is a whole different matter altogether (here is an interesting paper on the

work that Pangeanic did with Toshiba for EN<>JA). He also conceded that the output quality of statistical machine translation has almost reached a ceiling -- if it were not for the higher success through better domain-specific data.

(I agree that better training data will make the translation output better, but I also think that there is a ceiling that will stop the improvements. From a translator's perspective, I think that an even higher ceiling could be established if we had the TM, the termbase, and the MT systems "talk" to each other as I mentioned in the column linked to in the *Trados* article.)

So it's not surprising that *PangeaMT* users are encouraged to continue retraining their data with ever better materials to achieve better output. Right now this has to happen relatively manually: While a number of monolingual file formats are supported (TXT, DOCX, ODT, HTML) or are about to be supported (XLSX, PPTX, *OpenOffice* ODS/ODP, XML, CSV, *InDesign* IDML, *FrameMaker* MIF), translators typically pretranslate the supported bilingual files (TMX, TTX, XLIFF) in their translation environment tool, upload the files to the *PangeaMT* cloud to have the rest translated by MT, download the files, and post-edit them in their TEnT. Once the file is post-edited it is again uploaded as new training data for the MT.

Coming in September for several tools (presently *Wordbee*, *MemSouce,* and *XTM)* should be a more integrated process where the manual processes described in the last paragraph essentially would all be done tool-internally. There also will be an app in the *SDL OpenExchange* that will integrate *PangeaMT* into *Trados Studio* with a less thorough but still largely automated interaction with the MT cloud.

When I talked to Manuel about the system he insisted that it's a hybrid system (a combination of rules-based and statistical MT), and strictly speaking he's right because there are some rules about syntax that can be used (as described in the *PangeaMT*/Toshiba paper linked above). But since those seem the pretty much the only rules, I'm not sure I would call this a hybrid system.

But then that might simply be a matter of semantics. And isn't that it's about anyway!?

Excerpt from the 225th Tool Box newsletter by Jost Zetzsche