Using the Dutch Parallel Corpus to Calculate English-Dutch Word Translation Entropy

Large corpora are frequently used as linguistic data that is representative of a language, or at least representative of a specifically defined subset of that language. When a corpus comprises texts that are available in two or more languages, it becomes more than just a collection of translations. In addition to the language-specific properties of each translation, the relationship between these translations and especially between the different languages surfaces as well. These so-called parallel corpora allow researchers to compare languages empirically and allow them to make inferences about subjects of interest. The current study, situated within the PreDicT project (Predicting Difficulty in Translation)[1], consults such a parallel corpus, namely the Dutch Parallel Corpus or *DPC* (Macken, De Clercq, & Paulussen, 2011), to calculate word translation entropy for the English-Dutch language pair. DPC is managed and distributed by the CLARIN B Centre *Dutch Language Institute*, and available through CLARIN's Virtual Language Observatory[2].

PreDicT aims to build a system that, given an input text in language *x*, and a target language *y*, can predict how difficult it would be to translate said text to language *y*. On top of that, the system would highlight segments of the source text that are difficult to translate. Upon completion, the system will be made available through a public repository and accessible through a web-interface. The tool's metadata will be recorded in CLARIN's catalogs.

As a first step in creating this system, we completed a pilot study that correlated translation process data (duration, editing, and gaze features) with product data that, according to literature, can indicate translation difficulty (number of errors, word translation entropy, syntactic equivalence) (Vanroy, De Clercq, & Macken, 2018). The dataset that we used was taken from Daems (2016) and consists of a variety of translation process and product data collected by CASMACAT (Alabau et al., 2013) and Inputlog (Leijten & Van Waes, 2013) and merged by post-processing scripts. In total, there is information of 690 translated segments by 23 translators (13 professionals, 10 students). We found that all three product features indeed correlate with some process features, in particular with the number of times a translator has revised a segment's translation, and with the period of pause relative to the segment's total translation time.

In that pilot study, word translation entropy was calculated automatically as part of the process that converted the aforementioned dataset to the CRITT TPR-DB format (Carl & Schaeffer, 2018). To work out the entropy of a source word, the program looks up how it has been translated by all translators over all sentences. With a small translated corpus of 690 segments, entropy calculated in such a manner can be quite skewed and not representative. In the current study, we use DPC to get more justified entropy values. We use Moses (Koehn et al., 2007) to word-align the parallel corpus and retrieve the word translation entropy. Then, we calculate the correlations between our new-found entropy and the process data to compare these with the initial correlations from our pilot

[1] https://research.flw.ugent.be/en/projects/predict
[2] https://vlo.clarin.eu/

study. By doing so, we get an idea of how reliable it is to calculate entropy on small corpora that are based on different translations of the same segments.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., … Tsoukala, C. (2013). CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics*, *100*(1). https://doi.org/10.2478/pralin-2013-0016

Carl, M., & Schaeffer, M. (2018). The Development of the TPR-DB as Grounded Theory Method. *Translation, Cognition & Behavior*, *1*(1), 168–193. https://doi.org/10.1075/tcb.00008.car

Daems, J. (2016). *A Translation Robot for each Translator* (PhD Thesis). Ghent University, Ghent, Belgium.

Koehn, P., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., … Moran, C. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic. https://doi.org/10.3115/1557769.1557821

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta: Journal Des Traducteurs*, *56*(2), 374–390. https://doi.org/10.7202/1006182ar

Vanroy, B., De Clercq, O., & Macken, L. (2018). Predicting Difficulty in Translation: A Pilot Study. In *Proceedings of the 3rd Conference on Technological Innovation for Specialized Linguistic Domains*. Ghent, Belgium.