Workshop "Conditionals, Corpora, and Translation"
Maarten Bogaards

Heuristic Translation Mining and Distributional Analysis: Using parallel and non-parallel corpora side by side

Parallel corpora are a useful tool for linguistic research in various ways. One recent methodological application, called 'Translation Mining' (van der Klis et al. 2017), involves quantitatively comparing the conceptual range of crosslinguistically recurring formal categories (e.g. the Perfect in English, French and Spanish). But what if the linguistic category of interest is a clear part of one, but not all of the languages under study? An example of this scenario is viewpoint aspect in Mandarin, which is expressed by a set of specialized suffixes (e.g. *-zhe* 着 and *-le* 了), as opposed to Dutch, which has little aspect marking as part of its grammatical system.

In this talk, I show that Translation Mining and parallel corpora are also useful tools when linguistic inventories vary, albeit with a different methodological goal: to get an empirically founded overview of the way a conceptual category is expressed in a language in which the means of expression for that particular category are diffuse or unclear. In other words, in this application Translation Mining functions not primarily as a comparative method, but as a heuristic—which is why I propose the term 'Heuristic Translation Mining' (Bogaards 2019). Specifically, I use the 'durative' and 'resultative' marker *-zhe* 着 to investigate the expression of these aspectual notions in Dutch.

At the same time, I argue that the heuristic employment of parallel corpora needs to be supplemented by using a non-parallel corpus, as the former raises questions that only the latter can answer. While the Dutch translations of Mandarin sentences with *-zhe* 着 point to an interesting pattern in Dutch consisting of a posture verb and a past participle, the semantic and syntactic distribution of this pattern can only be examined by following up the parallel heuristic with a non-parallel corpus analysis. In this way, using parallel and non-parallel corpora side by side provides a powerful way of examining the formal manifestation of conceptual categories across languages.

References

Bogaards, Maarten (2019). 'A Mandarin Map for Dutch Durativity: Parallel text analysis as a heuristic for investigating aspectuality.' *Nederlandse Taalkunde/Dutch Linguistics* 24, 157-193.

van der Klis, Martijn, Bert Le Bruyn & Henriëtte de Swart (2017). 'Mapping the PERFECT via Translation Mining.' *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 497-502. http://www.aclweb.org/anthology/E17-2080.