Let's collaborate! Methods and tools for the construction of a complex translation and interpreting corpus

Adriano Ferraresi (University of Bologna), Silvia Bernardini (University of Bologna), Maja Miličević Petrović (University of Belgrade)

Keywords: Translation and Interpreting; European Parliament data; Parallel corpus; Corpus alignment; Collaborative corpus building.

This paper describes the steps involved in building a corpus of European Parliament (EP) proceedings. One may wonder if the task is worth the effort, given that the language samples included in the corpus are among those most readily available to the research community, and generally viewed as inexpensive to acquire and process. Yet the similarities between EPTIC, the *European Parliament Translation and Interpreting Corpus*, and better known and widely used parallel corpus resources such as the Europarl corpus (Koehn 2005) are limited to their data source. EPTIC is a small corpus but, we will argue, has features that make it unique, and justify the time and effort spent building it.

Initiated as an offshoot of EPIC (Russo et al. 2012), and originally including only Italian and English speeches from 2004 and their interpreted and translated versions (Bernardini et al. 2016), EPTIC is now growing through a community effort. The current version, developed jointly by teams at 5 European universities, includes sentence-aligned sub-corpora in 5 languages (English, French, Italian, Polish and Slovene), encompassing 10 language combinations: bidirectional alignments are available for source and target texts, as well as for texts in different modalities (spoken/written source texts, and translations/interpretations). The overall size of the corpus, which consists of 20 multi-aligned sub-corpora, is approximately 400,000 words.

After illustrating the possibilities offered by EPTIC as a source of multilingual, parallel, comparable and intermodal data for the study of translated and interpreted language, the paper discusses the major methodological and technical challenges involved in its construction. In particular, we focus on the processes and tools used for a) text-to-text alignment at sentence level, taking into account the large number of bi-directional alignment directions; b) text-to-video alignment, allowing corpus users to access time-aligned multimedia files of the original and interpreted EP speeches from concordances; c) annotation with linguistic and contextual information; and d) indexing for consultation by the public through the NoSketch Engine platform (Rychlý 2007). To streamline these procedures and facilitate collection of new data, an online platform has been developed. In the final part of the paper we will present such platform and discuss what we see as the most promising directions for corpus enlargement (e.g. balancing the corpus in terms of speeches delivered by native vs. non-native speakers, and in impromptu vs. read-out mode), with a view to encouraging new teams of researchers to join the EPTIC team.

(Word count: 396 words)

References

Bernardini, S., A. Ferraresi and M. Miličević. 2016. "From EPIC to EPTIC – Exploring Simplification in interpreting and translation from an intermodal perspective". *Target* 28 (1).

Koehn, P. 2005. "Europarl: A parallel corpus for statistical Machine Translation". In *Proceedings of the Machine Translation Summit X*, Phuket, Thailand. 79–86.

Russo, M., C. Bendazzoli, A. Sandrelli and N. Spinolo. 2012. "The European Parliament Interpreting Corpus (EPIC): Implementation and Developments." In Straniero Sergio, F. and Falbo, C. (eds). *Breaking Ground in Corpus- based Interpreting Studies*. Bern: Peter Lang. 35–90.

Rychlý, P. 2007. "Manatee/Bonito - A Modular Corpus Manager". In *Proceedings of the Workshop on Recent Advances in Slavonic NLP*, Masaryk University, Brno. 65–70.