How to use a nanocorpus. Enriching corpora of interpreting

Camille Collard, EQTIS, Ghent University

Bart Defrancq, EQTIS, Ghent University

Corpus-based research into interpreting is still in its infancy. The late Miriam Shlesinger warned the scholarly community of interpreting studies that, even though corpus-based research was much needed in their field to attain the necessary degree of generalization and empirical validity, it would nevertheless be quite a challenge to collect the amount of data such studies required (Shlesinger 1998). She proved right: efforts were undertaken in various places to collect corpora of interpreting, notably interpreting carried out at the European Parliament (Bologna, Poznan, Ghent *inter alia*), but the amounts of data are still very modest (typically around 250,000 tokens, including source and target texts). This seriously limits the kind of questions researchers can answer with regard to this special kind of language usage. Results of coarse-grained analyses focusing on highly frequent lexical items, e.g. type-token ratios, head lists, etc. are fairly reliable (Bernardini *et al.* 2015, Kajzer-Wietrzny 2015, Defrancq *et al.* 2015), but it is currently impossible to conduct analyses on the same scale as what is common practice in translation studies. On the other hand, as the research interests in the field are also quite particular, e.g. a strong focus on cognitive aspects of interpreting, enriching the corpus with specific metadata (speech rate, disfluencies, gender, time tags…) allows us to answer new questions in the field of cognitive science. In our presentation will show what the metadata can tell us about the Ear-Voice-Span of interpreters and introduce the use of the transcription and alignment tool EXMARaLDA Partitur-Editor.