Developing corpora for the visual modality: The case of sign language corpora

Much of corpus linguistics is dominated by studies of corpora centred around spoken languages. These corpora typically consist of written texts or, to a much smaller extent, transcripts of spoken interaction. There is a growing (but still small) number of corpora which strive to be multimodal (i.e., representative of spoken language as it is used in both the audio and visual modality, e.g., Adolphs & Carter 2013). This is in spite of the fact that much about human language can be learned when taking a multimodal perspective. In this presentation we will outline some of the major challenges in working with sign language corpora (using as an example the British Sign Language Corpus, Schembri et al., 2011). Unlike spoken languages, sign language data cannot be studied in text form as sign languages lack widely used writing systems or and there are no standard transcription/notation systems (i.e. nothing like the IPA for spoken languages). Addressing these challenges in setting up a corpus devoted primarily to language in the visual modality will help inform the field of corpus linguistics generally, particularly those seeking to develop non-text (monomodal or multimodal) corpora.