Measuring Variation in African Sign Languages: a Data Driven Approach

Manolis Fragkiadakis[1,2] and Victoria A.S. Nyst[1,2]
[1]Leiden University Centre for Linguistics, Leiden University, Netherlands
[2]Leiden Centre of Data Science Research Programme, Leiden University, Netherlands

Between 2007 and 2014, two large video corpora of West African sign languages have been compiled at Leiden University. The first sign language (SL) corpus contains a set of discourse data of around 30 hours and 15 signers of Adamorobe Sign Language (AdaSL), which emerged spontaneously in response to the high incidence of hereditary deafness in the village of Adamorobe, Ghana. The second sign Language corpus contains the results of a sign language survey in the Dogon area of Mali, notably in Berbey, close to Hombori.

The goal of this study is to measure the phonological and lexical variation of the lexical databases compiled from the annotations of these corpora. Analysis of cross-linguistic variation of SL lexica as documented in corpora is critical for historical-comparative studies of sign languages as well as for discerning patterns of variation. As our understanding of how sign languages are accumulated in language families is mainly based on historical information it needs to be informed by language-internal evidence. The absence of a shared history between these rural SLs enables us to compare variation between related and unrelated SLs, not influenced by contact. Currently, phonological and lexical variation is being processed by looking at individual realizations of different signs in videos or pictures or by comparing specific semantic fields. Our proposed methodology enables researchers to assess variation on a broader scale. The lexical databases contain the glosses (uniquely identifying spoken language words that by definition refer to a particular sign form) used to annotate the respective corpora, along with their phonological codings according to SignPhon's phonetic description guidelines. SignPhon's phonological units can be divided in up to 46 fields and meticulously encode the shape of the signs. In our project we only used 9 of them, namely: "Number of hands", "Handshape strong", "Location type", "Set of selected fingers", "Position of MCP joints", "Position of finger joints", "Spreading of fingers", "Thumb opposition" and "Thumb contacts fingers". In our comparison we used the Adamorobe, Berbey and NGT (sign language of the Netherlands) sign languages. NGT, being a historically unrelated language compared to the AdaSL and Berbey SL, serves as a benchmark. In total, there were 383 signs for Adamorbe SL, 91 for Berbey and 40 for NGT. To visualize the overall phonological variation of the sign languages we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) map. T-SNE is a popular methodology when it comes to the exploration of high-dimensional data and we utilized it to create clusters of signs that have similar features.

For the lexical variation we measured the Levenshtein distance of the common glossed signs. Each sign token is represented as a string of characters formulated by the concatenation of its phonological codings. Since all sign tokens are coded with the same number of parameters according to SignPhon, we measured only the necessary edits based on substitutions (i.e.

when parameters are not identical for a given pair of forms). Such method can provide an enhanced evaluation of the discrepancy degree of the signs.

The result of the T-SNE map hints that there might be different clusters of signs based on their phonological features. However, contrary to expectations and due to sparse and inconsistent data, it is hard to extract tangible conclusions. Additionally, common glosses between the annotated corpora are deficient and as a result there are no significant outcomes. This study is the first step towards enhancing our understanding of how sign language lexical and phonological variation can be assessed based on such data driven approaches. To further our research, we plan on adding more phonological codings to the rest of the lexical databases and utilize different distance metrics.