Sign language corpora are rather exotic within corpus linguistics as they deal with language data for languages having no written form and consequently no orthography. Much of the effort in current sign language corpus projects goes into segmentation and lemmatization as these need to be done manually while they are relatively straight-forward und in many cases automatic steps for other languages. The lack of large-scale lexical databases for sign languages implies that many decisions to be taken in these steps are preliminary and subject to later revision. Therefore, it is of utmost importance to always have access to the original data. We present our approach that takes these requirements into account and provides multiple views on the data in order to support data quality assurance even if independent double-transcription often is not an option due to the immense cost.

To illustrate the approach, we present data from the map task as used in the Dicta-Sign project that collected data from four sign languages based on the same elicitation setting. This example also demonstrates where one can expect some parallels between sign and spoken language corpora.