Use cases FCS workshop CPH

Minimal use case

Students of languages and linguistics can gain experience with corpora and (hopefully) insights into real language use through simple search tasks. While conducting such tasks, they also recognize that there is an international cooperating community sharing their data through a common infrastructure.

An example for such a task would be to search (German data) for the word “ja”, which is ambiguous regarding its function - it can be used e.g. as a response (“yes”), a back-channel (“yeah”, “uhum”), a discourse particle (“well”) or as a modal particle (~ my proposition is a known fact on which we agree). By examining syntactic and phonetic features of the retrieved keywords, students can postulate rules for the forms corresponding to the various functions, approaching the subject inductively. The students could also compare features and functions of the word “ja” in written vs. spoken language.

This kind of inductive approach does not require any specific common tagset or annotation scheme to be considered as the search is conducted on the transcribed text. However, since “literary transcription” is often used to transcribe conversations, a straightforward way to also retrieve variants, e.g. as described by “[Jj]a+” would facilitate searching. Apart from the retrieved keywords, the conversational context with corresponding recorded audio is sufficient as input to conduct a simple analysis. Efficient searches would require metadata on the availability of recordings to restrict the search at an early stage, and ideally also on the alignment status, since poorly aligned transcriptions might be of little use in this case.

Though recordings and original transcription files containing the required context are available through PIDs resolving to CMDI or directly to content files, the user would be confronted with a plethora of formats. The KWIC view for the search result displays the context of written text and single speaker turns appropriately, but for conversational data, surrounding speaker turns are necessary. Using information on time, tiers and speakers, an audio-aligned visualisation based on commonly used transcript layouts - lists, scores (or tables) - could be generated for each search result, provided a common exchange format for multi-tier transcription data.

More elaborate use case

A researcher in the field of interlanguage has encountered an interesting phenomenon regarding intonation in yes/no-questions in her corpus of spoken learner language from Japanese learners of German. The researcher now wishes to asses the scope of her findings by comparing the data with that of learners with different L1s. By defining the required meta data on speakers’ language knowledge, the researcher can restrict the search for sentences succeeded by a speaker change with a contribution containing either “ja” (“yes”), “nein” (“no”) or “doch” (“yes” in reply to negative questions/sentences) to relevant resources. Since many of the resources require personalized access (RES), search results from resources to which the researcher has no access are only displayed as placeholders with information on how to request access. The researcher can browse the search results and e.g. remove false positives. For the analysis, the simple transcription view and the aligned recordings would allow for a comparison of the intonational features, but the researcher might also decide to collect the relevant data and annotate it automatically and manually to support the analysis.

Apart from the basic support for spoken conversational data outlined in the simple use case, the kind of search conducted in this use case would require more sophisticated search functionality and access to speaker metadata. An interesting question is what happens after the search has been conducted - how can the results be further processed and analysed?