A Client-Side Web Application for Advanced Searching of Linguistic Corpora

This presentation discusses the planning and implementation of a client-side powered web-based application for quick advanced searches of linguistic corpora for aid in language documentation, originally developed for the purpose of advanced queries on a FLEx based corpus of Ende (Pahoturi River). The application consists of a query language and an interpreter which takes in a JSON file representing the corpus as input. I discuss the organization of the file format and the structure of the query language as well as how corpora such as those using FLEx, LaTeX, or plaintext files may be processed to fit such a JSON file format to be used for searches.

The JSON based file format structures corpus materials into three components: a phrase dictionary, word dictionary, and morph dictionary. The phrase dictionary contains a set of all phrases in the corpus, each with a 'words' attribute whose value is defined as an array of IDs for each word in the sentence. The word dictionary contains a set of all words in the corpus with defined IDs and a 'morph' attribute whose value is defined as an array of IDs for each morph(eme) in the word. Additionally, entries of each dictionary may contain user defined attributes, such as translations, notes, glosses, and tags. The benefit of such a file structure where individual words and morphemes are represented only once is that file size can be much smaller than the original corpus from which the JSON file is derived without loss of information, allowing faster and less cumbersome file uploading and manipulation.

The query language resembles that of SQL, in which tokens are divided into keywords, strings, and numbers, but is run client-side via javascript. Various functions are defined which allow a user to perform advanced filters on the corpus, as well as, manipulate data, display data as two- or three-lined glosses via html, and download subsets of the corpus as a csv file. Emphasis is made on ease of use for the query language such that people inexperienced with coding or linguistics may be able to search corpora with ease. For use with FLEx, the FLEx corpus data is first downloaded into .flextext files, which are converted to ready-to-use JSON via python code. Other methods for conversion from LaTeX and plaintext based corpora are also discussed. All materials are open source and accessible via Github repository.