:: The Taa language corpus ::

The structure of the Taa language corpus is intended to help locate and identify relevant resources both according to formal criteria and with regard to content. The most paramount distinction is between 'Primary data', i.e. recordings and other media-files, and 'Secondary data' such as linguistic and anthropological analysis, summaries, publications and other products. (The node 'Taa Team Only' contains internal files which are not accessible for external users.)

The next level of distinction is that between the two countries where Taa is spoken. They are kept separate because of the different level of analysis of the recordings. While the two varieties spoken in Namibia have been subject to an in-depth research for several years, the varieties spoken in Botswana could so far only be approached by means of short-term surveys. The primary data are further distinguished into non-speech recordings (photographs, non-speech videos, drawings) on the one hand and speech recordings on the other.

The Namibian 'Speech'-branch distinguishes between language varieties ('N|oha, West !Xoon, etc.), while the Botswanan 'Speech'-branch is arranged according to the villages where the recordings were made because, for Botswana, the analysis of dialect borders is not yet finalized. The next distinction is between 'elicitations' (lexicon, grammar, phonology, etc.) and 'texts' (conversations, descriptive texts, narratives, etc.). Further down the tree the 'texts' are arranged according to topics.

The session names are standardized. All session names start with the letter 'T' for Taa. The second letter identifies the language variety or speech community, i.e. 'N' for 'N|oha, 'W' for West !Xoon and 'T' for so far unidentified varieties or ambiguous assignments. For Taa in Botswana, three letters, usually the first three letters of a place name, indicate the respective villages. The small letter denotes types of modality: 'a' stands for natural speech (texts in vernacular language such as conversations, monologues, etc.), 'b' is used for elicited speech, 'c' for speech in vehicular language (translations or content descriptions, interviews in a vehicular language, etc.) and 'd' for no speech (no substantial linguistic/language information). These letters are followed by the date of the recording in a 6-character-format indicating the year, month and date, and an ordinal number for different recordings of the same type made on the same day.