Not last, even if least: endangered Formosan aboriginal languages and the corpus revolution

Dr. Josef Szakos and Amy Wang
English Department, Providence University, Taichung, Taiwan

In our paper we intend to report on the present stages and uses of corpus creation for the Austronesian languages of Taiwan. Our main attention goes to the Tsouic tribes, including Northern Tsou, Kanakanavu and Sa' arua. The latter two can be regarded as highly endangered as there are only dozens of old speakers left. Our corpus serves a three-fold purpose: Authentic documentation of the oldest remnants of early Austronesian languages, making them analysable for the linguistic community around the Globe, and enabling a revitalisation and stabilisation of language use in the aboriginal communities. As these languages have lacked writing systems until now, the concomitant task of alphabetisation has to be solved, too. The dozens of hours of transcribed speech data are searchable by any kwic programs and concordancers. The innovation of our approach is the linking of sound files (the length of intonational units) with the corpus, and consequently with the output of searches. This makes not only the instruction of speech patterns for young people easier, but it also provides the researchers with the phonetic context. The quick comparison of intonation patterns can give additional information for semantic subdistinctions. The corpus is constantly growing, providing a series of CD-s, while we also intend to combine it with web-availability. For the young learners of their mother-tongue, we have created a sympathetic user interface including a combined vocabulary. Remaining problems are the possible automatization or partial automatization of the recording and segmentation procedure, search for topics, themes and balancing the corpus by including the speech of more persons of the last speakers. Since the users, the mother-tongue learners and researchers need an interface in Chinese or English, we still have to work out better solutions for that (like parallel corpus arrangement). We hope that the overview and introduction of some problems may raise the attention of some experts who could solve the further problems.