Cultures and Corpora: Extracting Anthropological Information from Corpora of Formosan Endangered Languages

Dr. Jozsef Szakos
Department of English Language, Literature, and Linguistics
Providence University, Taichung, Taiwan
szakosdr@ms38.hinet.net

Corpus research has concentrated on finding grammatically relevant information in collections of well-documented languages. While applying the methods of CL to minority languages, we have encountered another area worthy of discussion, namely the treatment of cultural and anthropological information in these corpora.

For English, French or other major corpus research areas the solution has been either to ignore cultural issues influencing grammar, or to relocate such information into encyclopaedias, which in turn were not regarded as corpora. To illustrate this point, I just mention that Encyclopaedia Britannica is not BNC, but BNC cannot replace the knowledge contained in Britannica. At the same time, we are conscious that the relevant information in our culture also influences the grammar structures in BNC.

While we are producing corpora for endangered languages, which I intend to exemplify on Austronesian languages of Taiwan (Formosan languages), we cannot escape two problem-fields. The first one is that we need to create bilingual corpora. The second necessity is the inclusion and treatment of anthropological and cultural information in these collections. We are going to deal with about a dozen languages, so it is desirable to work out the standards of bilingualor plurilingual corpora, and at this point we encounter the problem ofethnologically relevant information and its extraction.

In my paper I intend to discuss the above questions while introducing the present stage of our research. I wish to speak about naming and culture (including conceptualisation of living and non living things), deixis and culture (geography, social life determining factors), language use and culture (constraints on the use of some constructions). This would also include analysing the existence or non-existence of ethnically conditioned concepts. I will also demonstrate some technical solutions for the bilingual corpora, using ParaConc. Besides, I want to show how this approach of taking into consideration the local cultural phenomena can contribute to the creation of authentic teaching materials for the preservation of these languages.