Spoken Subcorpus of the Kazakh Language: History, Content, Methodology

Authors : Kuralay Bimoldaevna Kuderinova, Beisenkhan Samal
Abstract : The history of creating a linguistic corpus in Kazakh linguistics begins only in 2016. Though within this short period of time, the linguistic corpus has become a national corpus and its several subcorpora, namely historical, cultural, spoken, dialectological, writers' subcorpus, proverbs subcorpus and poetic texts subcorpus, have appeared and are working effectively. Among them, the spoken corpus has its own characteristics. The Kazakh language is one of the languages belonging to the Kypchak-Nogai group of Turkic peoples. The Kazakh language is a language that, as a part of the former Soviet Union, was directly influenced by the Russian language and underwent major changes in its spoken and written forms. After the Republic of Kazakhstan gained independence, the Kazakh language received the status of the state language in 1991. However, today, the prestige of the Russian language is still higher than that of the Kazakh language. Therefore, the direct influence of the Russian language on the structure, style, and vocabulary of the Kazakh language continues. In particular, it can be said that the national practice of the spoken language is disappearing, as the spoken form of Kazakh is not used in official gatherings and events of state importance. In this regard, it is very important to collect and preserve examples of spoken language. Recording exemplary spoken texts, converting them into written form, and providing their audio along with orphoepic explanations will serve as a valuable tool for teaching and learning the Kazakh language. Therefore, the report will cover interesting aspects and scientific foundations related to the creation, content, and methodology of the oral subcorpus of the Kazakh language.
Keywords : spoken corpus, Kazakh language, orthoepic norm, LLM
Conference Title : ICACL 2025 : International Conference on Applied Corpus Linguistics
Conference Location : Amsterdam, Netherlands
Conference Dates : May 13-14, 2025