World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:11, No:04, 2017

Grammatically Coded Corpus of Spoken Lithuanian: Methodology and Development

Authors : L. Kamandulytė-Merfeldienė

Abstract : The paper deals with the main issues of methodology of the Corpus of Spoken Lithuanian which was started to be developed in 2006. At present, the corpus consists of 300,000 grammatically annotated word forms. The creation of the corpus consists of three main stages: collecting the data, the transcription of the recorded data, and the grammatical annotation. Collecting the data was based on the principles of balance and naturality. The recorded speech was transcribed according to the CHAT requirements of CHILDES. The transcripts were double-checked and annotated grammatically using CHILDES. The development of the Corpus of Spoken Lithuanian has led to the constant increase in studies on spontaneous communication, and various papers have dealt with a distribution of parts of speech, use of different grammatical forms, variation of inflectional paradigms, distribution of fillers, syntactic functions of adjectives, the mean length of utterances.

Keywords : CHILDES, corpus of spoken Lithuanian, grammatical annotation, grammatical disambiguation, lexicon, Lithuanian
Conference Title : ICCLM 2017 : International Conference on Corpus Linguistics and Methodology
Conference Location : Venice, Italy
Conference Dates : April 13-14, 2017