DEBRECENI EGYETEM
FINNUGOR NYELVTUDOMÁNYI TANSZÉK
FOLIA URALICA DEBRECENIENSIA 30.
DEBRECEN, 2023

PermCorp: Towards the implementation of a Komi-Permyak corpus

In this paper we present the structure and the process of creating a new corpus of the Komi-Permyak language. The aim of the project is to collect the available literature in Permyak and to provide glossing and English translation for the texts. Since Komi-Permyak is an under-researched language, it is extremely important to produce a linguistically usable text collection that contributes to making the language more visible. The primary purpose of PermCorp is to represent the written version of the Permyak language in public use. In order to do this, we have collected texts of different genres, each of which will be represented by a sub corpus, and which have been balanced by text type. After the digitalization of the printed works and the transliteration, the texts will be labelled with multi-level tagging (glossing, POS tagging, English translation). During the project, we use the FLEx software, because its morphological analysis function with the dictionary and grammatical outline could help the researchers effectively in tagging.

Keywords: corpus, Komi-Permyak, FLEx, POS tagging, English translation

SZILVIA NÉMETH – DITTA SZABÓ – NIKOLETT F. GULYÁS