Nikolett F. Gulyás[1] – Szilvia Németh[1] – Ditta Szabó[1,2]
([1]Eötvös Loránd University, [2]HUN-REN Hungarian Research Centre for Linguistics)

Preparing the FLEx morphological analyzer for the Komi-Permyak language

Within the project "Komi-Permyak corpus" (NKFIH FK 143242), our research group is working on the construction of a 300,000 token Komi-Permyak corpus (PermCorp) containing written texts of several periods, pre-labelled the grammatical annotations with an automatic tool and then finalized by manual checking, with English sentence translations (Németh et. al 2023). Although the project has to deal with the challenges that arise when building a diachronic corpus of almost any minority language in Russia with a small number of speakers (from estimating the frequency of text types available in the language to writing systems that change several times over time to programming a user-friendly search interface), in this talk, we will present the problems we encountered when preparing a general-purpose, fixed-structure morphological analyzer for machine processing of an under-documented, incompletely described language.

Since the aim of our project, besides the development of the corpus, is to publish a morphological analyzer that we use and that can be easily used by other researchers, independently of the level of their competency in Komi-Permyak, we have decided that rather than creating our own morphological analyzer, we will apply a general analyzer for the Komi-Permyak language system which already has a user interface and user documentation. So that others can use this tool to machine-generate the morphological analysis of the data they have collected or that are not in the corpus, published in written form. Our choice was FieldWorks Language Explorer (FLEx), one of the few documentation software that is suitable for processing purely written texts and also has a general purpose analyzer.

In corpus building, authors usually create their own morphological analyzer, which provides a high degree of flexibility, but PermCorp's authors decided on adapting to the structure of the FLEx analyzer and accept its limitations. The FLEx general analyzer allows the user to define phonological and morphological rules according to a predefined set of rules known in the language, so that it can efficiently segment word forms into morphemes using the added dictionary (the analyser does not deal with context, word environment). The resulting Permyak morphological analyzer and dictionary will be published as a FLEx file (‘project’), so that we can provide an immediately usable tool for other researchers wishing to annotate.

In this talk, we will illustrate what it means to adapt to the FLEx analyzer in a language where, for example, morphotactic rules are not described in sufficient depth, by means of adjective and verb categories. We would like to show that technical constraints sometimes influence the choice of the main glossing principles (e.g. for verb mode marking or for adjectives functioning occasionally as nouns), while at other times a multistage rule system needs to be established (e.g. to introduce a category that treats nouns and adjectives as a unit or to handle morphological alternations like [v] ~ [l]). Some issues require a separate corpus

analysis in the absence of exhaustive language descriptions (e.g. morphotactic rules for adjective suffixes), while for others we have not found a way to use the FLEx automatic analyzer to produce a clear output without human judgment (e.g. in case of verb syncretism). We also present our experiences with the FLEx parser's ability to handle large token counts, its support for collaboration between multiple linguists, and the slowdown of the software due to the combination of lexicon and grammar rules.

References

Németh, Szilvia – Szabó, Ditta – F. Gulyás, Nikolett 2023. PermCorp: egy komi-permják korpusz létrehozása. [PermCorp: developing a Komi-Permyak corpus]. *Folia Uralica Debreceniensia* 30: 181–202. https://real-j.mtak.hu/26994/1/fud30.pdf