Project title:

Expanding the Ukrainian UD treebank and exploring the possibilities of using it for annotation of large corpora and in teaching.

Name and affiliation:

Maria Shvedova, Ukraine, National Technical University «Kharkiv Polytechnic Institute», Department of Intelligent Computer Systems, lecturer.

Keywords:

Ukrainian, corpus annotation, treebank.

Description:

Currently, there is only one UD treebank for the Ukrainian language, that has not been expanded since 2019 and has not very high accuracy for morphology compared to the VESUM dictionary-based morphological analyzer that we most often use to annotate Ukrainian corpora [1].

The existing resources relevant for the present application are the GRAC corpus [2], developed since 2016, and the recently inaugurated family of parallel Ukrainian corpora ParaRook. These are corpora of standard-oriented written Ukrainian, featuring, in the case of GRAC, the information of the territory related to the text and to the author.

Although I have no experience in creating UD corpora, I am motivated to learn and work on improving UD for Ukrainian for the following reasons:

1. The need to add UD tagging to the large Ukrainian language reference corpus GRAC as extra layer of annotation. This task is in its planning stage. This would be a significant improvement for GRAC, as it currently lacks a syntactic annotation at all. Also, GRAC is used by linguists from different countries for whom the UD system is more familiar than the one we use now.

2. The Ukrainian parallel corpora (ParaRook) are in the initial development phase, the first one, the German-Ukrainian parallel corpus, was recently published [3]. For tagging parallel corpora, we already use the authomatical UDPipe2 annotation, which is convenient for multilingual resources. However, the accuracy for the Ukrainian language could be improved. The current accuracy of UDPipe2 for morphology is 91% (the VESUM-based morphological analysis system provides 99% accuracy for modern standard texts).

3. My students show great interest in UD and want to develop a corpus for Ukrainian. I would like to explore the possibilities of involving students in corpus

annotation and adding this activity to the corpus linguistics course that I teach. Two of my Kharkiv students are also applying to participate in this workshop (we ask for the possibility of online participation for them, as men of their age are not allowed to leave Ukraine at the moment due to ongoing war).

Among the open questions to be addressed during the hackathon one might cite annotation of a syntactic golden standard for Ukrainian and possible issues related to POS disambiguation in this context.

References.

1. V. Starko, A. Rysin, VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In: Computational Linguistics and Intelligent Systems. Proc. 6th Int. Conf. COLINS 2022. Volume I: Main Conference. Gliwice, Poland, May 12–13, 2022, pp. 71–80.

2. Maria Shvedova, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko et al. (2017-2024): GRAC: General Regionally Annotated Corpus of Ukrainian. Kyiv, Lviv, Jena. Available at uacorpus.org.

3. Maria Shvedova, Arseniy Lukashevsky. ParaRook||DE-UK - parallel German-Ukrainian and Ukrainian-German corpus based on GRAC. / NTU KhPI, Department of Intelligent Computer Systems. Available at https://uacorpus.org/Kyiv/pararook