## A Contemporary News Corpus for Ukrainian (CNC-UA)

**S. Fischer, K. Haidarzhyi, J. Knappen, Y. Stodolinska & E. Teich**
*Universität des Saarlandes*
stefan.fischer@uni-saarland.de, kateryna.haidarzhyi@uni-saarland.de,
j.knappen@mx.uni-saarland.de, yuliya.stodolinska@uni-saarland.de,
e.teich@mx.uni-saarland.de

We present a corpus of Ukrainian contemporary news articles (CNC-UA) comprising 87,210,364 words and 292,955 texts. The sources represent standard language and were published between 2019 and 2022 on https://suspilne.media, the news website of the national public broadcaster of Ukraine. The motivation for building the corpus was to track language use in news reporting as the Russian war against Ukraine proceeded. We will show selected analyses at the poster.

While the number of Ukrainian speakers is around 45 million people, Ukrainian can still be considered a low-resource language due to the limited availability of non-commercial resources for language processing and research and the scarcity of publicly available corpora. To our knowledge, there is no similar corpus of recent Ukrainian news articles from a single source that is available for research.

CNC-UA was built from a database dump provided to us by Suspilne. Linguistic annotations were added by processing the texts with the Stanza NLP library (Qi et al., 2020). Each text is annotated with an identifier, article title as well as date and time of publication. Currently, we apply various language modelling techniques to the corpus, including topic models, for analysis of the data.

The corpus is available for non-commercial use. We provide two tab-separated formats: CoNLL-U from the Universal Dependencies project (de Marneffe et al., 2021) and vertical text format (VRT) as used by the CWB (Evert & Hardie, 2011) and CQPweb (Hardie, 2012). The corpus is hosted at the Saarbrücken CLARIN center (hdl:21.11119/0000-000E-1C5C-D) under a CC BY-NC-ND licence.

The authors acknowledge financial support from Deutsche Forschungsgemeinschaft (DFG) – project IDs 460033370 (Text+) and 232722074 (SFB 1102) as well as the Federal Republic of Germany and the 16 federal states in the framework of the National Research Data Infrastructure (NFDI) and its association NFDI e.V.

**References:** • Evert, S. & A. Hardie (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. University of Birmingham, UK. • Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380–409. • de Marneffe, M.-C., C. D. Manning, J. Nivre & D. Zeman (2021). Universal Dependencies. *Computational Linguistics* 47(2), 255–308. • Qi, P., Y. Zhang, Y. Zhang, J. Bolton & C. D. Manning (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online.