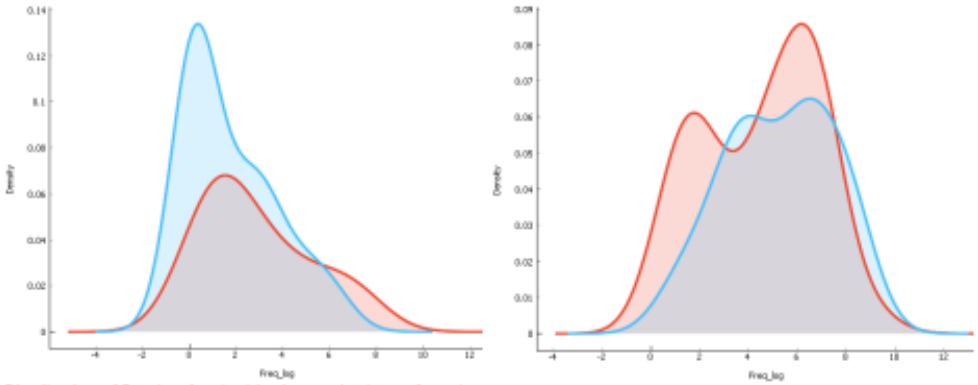
Exploring (mor)phonotactic variations: Word-Initial Consonant Clusters in Ukrainian

Keywords: Ukrainian, phonotactics, morphonotactics, consonant clusters, the Slavic languages

This study aims to analyze the phonotactics and morphonotactics of Ukrainian, comparing them qualitatively and quantitatively and explaining the differences between these two perspectives. A corpus-based approach is utilized to quantitatively analyze the morphological composition, preferability, and frequency of consonant clusters in Ukrainian. The study relies on the extensive electronic corpus GRAC (Shvedova et al., 2017) to make initial generalizations about the distribution of morphonotactic and lexical patterns of Ukrainian consonant clusters.

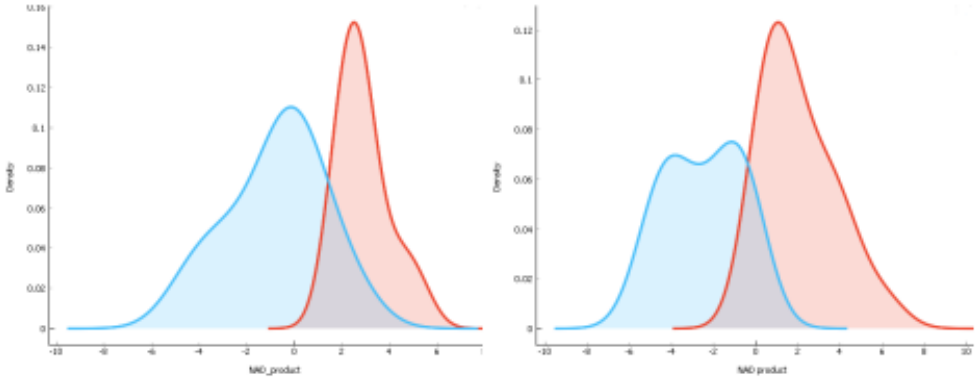
Two hypotheses were tested in this study. Hypothesis 1, previously formulated by Dressler & Dziubalska-Kołaczyk (2006), suggests that the degree of phonological preferability is inversely proportional to morphological complexity. Thus, morphonotactic consonant clusters are expected to be less preferred than phonotactic ones. The second hypothesis states that the degree of cluster preferability is directly proportional to frequency. Preferred clusters are expected to be more frequent than dispreferred. The NAD calculator (Dziubalska-Kołaczyk 2007) was used to determine the status of peripheral consonant clusters. Thus, the analysis shows that among 112 word-initial double consonant clusters, 61 are preferred, and 51 are dispreferred. The majority of word-initial double clusters were phonotactic, providing strong support for Hypothesis 1. For triple consonant clusters, the majority were morphonotactic and strongly dispreferred (N=41) vs. preferred (N=28), aligning with Hypothesis 1.

The second hypothesis was tested using statistical analysis performed in Orange, an open-source data mining toolbox for Python (Demsar et al., 2013). Linear regression was employed to examine the relationship between selected variables, particularly the NAD Product and frequency per million (FreqMil). Due to some outliers in FreqMil, a logarithmic transformation (FreqLog) was applied before conducting linear regression analysis (Fig.1).



Distribution of FreqLog for double clusters (red = preferred; blue = dispreferred).

Distribution of FreqLog for triple clusters (red = preferred; blue = dispreferred).



Distribution of NAD Product for double clusters (red = preferred; blue = dispreferred).

Distribution of NAD Product for triple clusters (red = preferred; blue = dispreferred).

***Figure 1**. Distribution FreqLog and NAD Product for double and triple clusters.*

In conclusion, the study confirmed the general presumption that morphonotactic clusters tend to be marked and dispreferred. However, the statistical analysis revealed only a weak correlation between the frequency of consonant clusters and their preference according to the NAD.

Word-count: 430

References

Dressler, Wolfgang U. and Katarzyna Dziubalska-Kołaczyk. 2006. “Proposing morphonotactics”. *Rivista di Linguistica* 18(2): 249-266.

Demsar Tomaz, Curk, Ales Erjavec, Crt Gorup, Tomaz Hocevar, Mitar Milutinovic, Martin Mozina, Matija Polajnar, Marko Toplak, Anze Staric, Miha Stajdohar, Lan Umek, Lan Zagar, Jure Zbontar, Marinka Zitnik & Blaz Zupan 2013. “Orange: Data Mining Toolbox in Python”, *Journal of Machine Learning Research* 14. 2349−2353.

Dziubalska-Kołaczyk Katarzyna 2014. “Explaining phonotactics using NAD”. *Language Sciences* 46, part A, 6–17.

Shvedova, Maria, Ruprecht von Waldenfels, Sergiy Yarygin, Andriy Rysin, Vasyl Starko & Tymofij Nikolajenko et al. 2017-2022. *GRAC: General Regionally Annotated Corpus of Ukrainian*. Electronic resource: Kyiv, Lviv, Jena. Available at uacorpus.org.