A Ukrainian-Russian Code-Switching Corpus of Ukrainian Parliamentary Sessions Transcripts

Maria Shvedova [*†] **Olha Kanishcheva**[*]
[*]University of Jena
[†]National University "Lviv Polytechnic"

Relevant UniDive working groups: WG1

1 Introduction

In this paper, we present the Ukrainian-Russian Code-Switching Corpus of Ukrainian Parliamentary Session Transcripts (1990-2020), its composition, annotation, and research possibilities. The language markup in the corpus is carried out at the sentence level, and the percentage of Ukrainian is determined for sentences using two languages.

The corpus represents bilingual Ukrainian-Russian parliamentary discourse, which has been changing over the years and became monolingual Ukrainian in the second half of the 2010s. The corpus provides an opportunity to analyze the actual use of Ukrainian, Russian, or both languages for each speaker and party, within a year or convocation. This helps to trace the connection between language use and the political position of the speaker/party, as well as trends in language use in the parliament and the political situation.

As a result of Ukraine's long history of political dependence on Russia, a significant number of people in Ukraine are bilingual in Ukrainian and Russian. Since Ukraine's independence (after 1991), the share of the use of the Ukrainian language in society has gradually increased and the share of Russian has decreased; the war of 2022 has significantly accelerated this process Kulyk (2022).

The Russian-Ukrainian bilingualism is characterized by code-switching that was also prominent in parliamentary speeches. Creating a corpus is a promising method of studying code-switching, as it allows us to see code-switching in a broader linguistic context and quantify language use. A typology of code-switching found in parliamentary speeches is to be presented in a separate section.

The experience of compiling code-switching corpora based on parliamentary texts already exists: these are bilingual Dutch-French speeches from the Belgium Federal texts Marx and Schuth (2010) and the Bilingual Corpus of Basque Parliamentary Transcriptions Escribano et al. (2022). For example, BasqueParl shows that there has been no significant change in the amount of bilingualism in parliament over the period 2012-2020, which is covered by the corpus [p. 3387]. The specific feature of the Ukrainian corpus of parliamentary transcripts from 1990-2020 is that the proportions and use of the two languages in it change noticeably and unevenly over the years, from the lowest share of Ukrainian at 76% in 1995 gradually reaching 100% Ukrainian in the second half of the 2010s.

Number of files	1957
Number of sentences	826 471
Number of tokens	16 657 948

Table 1: The quantitative data of our corpus.

2 Corpus Description

The corpus of the Verkhovna Rada (the Ukrainian unicameral parliament) proceedings contains texts recorded from 1990 until 2020, downloaded from the official website of the Verkhovna Rada. The timespan starts even before Ukrainian independence when Verkhovna Rada was an institution of a Soviet republic. The size of the corpus is about 70 million tokens. General information on these files is 290 presented in Table 1.

The corpus consists of text files, each of which contains all the transcripts for a year. The parliamentary speeches and remarks are recorded literally, in the language actually spoken, and language mixing is also accurately reproduced. This accuracy allows us to analyze the use of a particular language in a dialogue, depending on the language of the other interlocutors and the topic of the session.

The corpus also exists in another version, organized by speaker, that is, as text files that contain all the speeches of each deputy for the year (not unlike the Hansard website of British parliamentary discussions). This form allows us to analyze the use of Ukrainian and Russian by each speaker. The corpus is annotated by age and the political party of a given speaker, as well as by the adminis-

trative region of Ukraine (or by another country if applicable) where they were born and studied.

It is necessary to determine the speech language of each speaker. The sentence was chosen as the marking unit. The existing Lingua-py library was used to determine the language. It works accurately save for very short sentences that are often orthographically ambiguous between Ukrainian and Russian.

3 Types of Code-Switching

In our data, we observed the following most common patterns of mixing Ukrainian and Russian in parliamentarians' speeches:

- Ukrainian speakers insert phraseology or quotations in Russian.
- Russian speakers insert the names of laws and documents in Ukrainian.
- Russian speakers insert Ukrainian words and language clichés.
- Ukrainian speakers insert Russian words.
- Unmotivated heavy mixing of Russian and Ukrainian (Surzhyk);
- The language distinguishes between the official position proclaimed in Ukrainian (possibly read from notes) and personal opinions added in Russian;
- Triggered code-switching.
- Another language marking quoted speech.
- Switching to another language to illustrate a tolerant attitude to linguistic diversity.

4 Data analysis

We split up the texts into single files per speaker and year for the parliamentary sessions from 2010 to 2019. The data were grouped by convocations. For each convocation, the number of speakers who use Ukrainian, Russian, or both languages were counted.

A quantitative report of parliament speeches by language for each year (1990-2021) is given in Figure 1. The diagram shows that the share of the Ukrainian language in the corpus is gradually increasing and reached 100% in 2018.

This has been influenced by a combination of policy changes and relevant legislation passed over the years. However, none of these judicial acts seem to have had a discernable impact on language use in the parliament, as we had initially supposed. Rather, the use of Ukrainian consistently increased over the period from 2007 through 2014, before stagnating and finally growing again in 2017/2018.

We can assume the influence of political trends on the language in some cases (e.g. 2007, when an increase in the share of the Russian language coincided with the pro-Russian campaign in the Rada) on the degree of bilingualism in the Verkhovna Rada, but this needs additional research.

In the future, we plan to process the entire corpus of parliamentary transcripts for 1990-2020 and consistently trace the manifestations of Ukrainian-Russian bilingualism over 30 years and the history of its fading. We found some typical cases of bilingual speeches on the material of 2003 texts, and we want to look for similar cases automatically and trace the trends of different cases (language mixing and language switching) in the Rada over the years. In the future, additional corpus labeling is planned, such as part of speech, and entities will make it possible to identify additional connections between speakers. It would be interesting also to apply thematic modeling and trace the correlation between the discussion of the language issue in parliament and the actual use of languages. The approach taken in this paper can be extended to the corpus-based study of parliamentary transcripts of institutions where language diversity, including code-switching, is manifested, including, but not limited to, the diachronic dimension. The institutions in question include parliaments of multilingual territorial entities as well as councils of international organizations with multiple working languages.

Acknowledgments

We would like to thank Kyrylo Zakharov for downloading the plenary session transcripts from the Verkhovna Rada website. This research was partially funded by the Humboldt Foundation and the Volkswagen Foundation.

References

Nayla Escribano, Jon Ander Gonzalez, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez-de Viñaspre, and Rodrigo

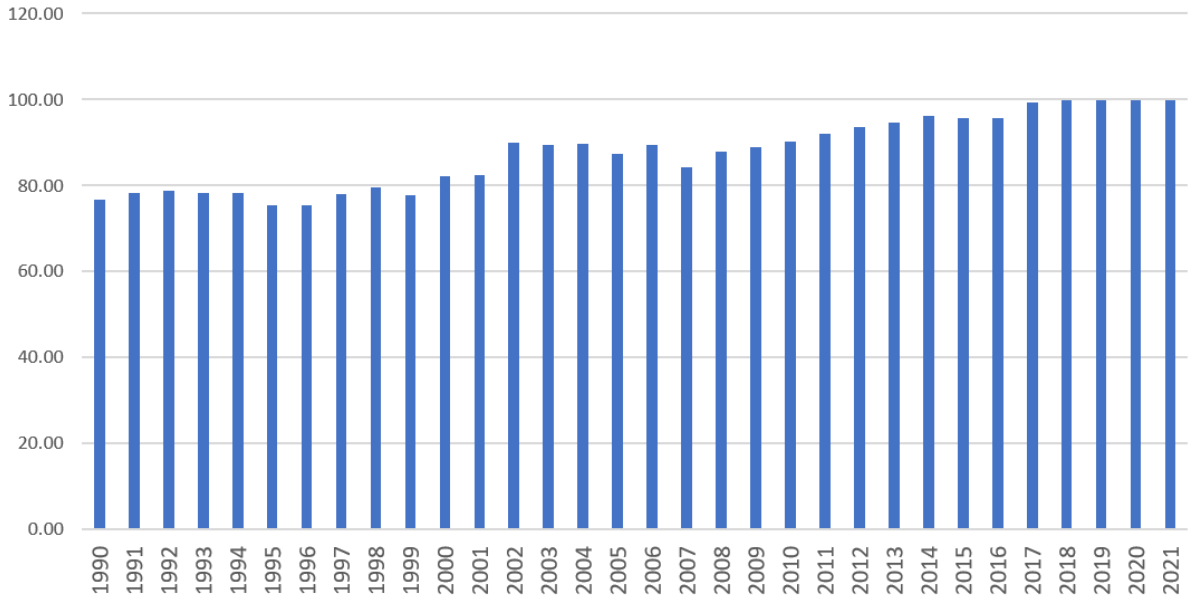

Figure 1: A quantitative report of parliament speeches by language for each year (1990-2021).

Agerri. 2022. BasqueParl: A bilingual corpus of Basque parliamentary transcriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3382–3390, Marseille, France. European Language Resources Association.

Volodymyr Kulyk. 2022. Language and identity in ukraine at the end of 2022.

Maarten Marx and Anne Schuth. 2010. DutchParl. the parliamentary documents in Dutch. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).