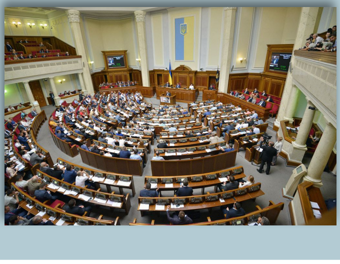
A Ukrainian-Russian Code-Switching Corpus of Ukrainian Parliamentary Sessions Transcripts

Maria Shvedova (Lviv Polytechnic, University of Jena),
Olha Kanishcheva (University of Jena)
WG1

Abstract

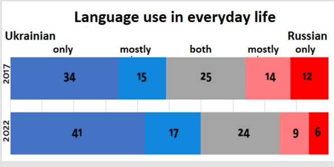
The paper presents the Ukrainian-Russian code-switching corpus of Ukrainian Parliamentary Session Transcripts. It features not only code-switching texts but also the entire bilingual discourse. It includes speeches of politicians who use Ukrainian, Russian, or various types of mixed speech and switch between these depending on the communicative situation.

The paper shows the process of creating this corpus from the original transcripts. Judging by the obtained information about the language choice, the reasons behind the variation in the degree of parliamentary bilingualism are considered.



Introduction

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 issues of Ukrainian-Russian bilingualism have been the subject of many studies by Ukrainian linguists in the 21st century (mainly in the context of sociolinguistic studies of the spread of both languages and the tasks of supporting the Ukrainian language and preventing the mixed Ukrainian-Russian speech known as Surzhyk).

Corpus-based studies of Ukrainian-Russian bilingualism have not yet become widespread. An exception is the Oldenburg Surzhyk corpus, which consists of mixed speech recordings made by researchers in different regions of Ukraine, and the studies based on it that examine the distribution of different variants within mixed Ukrainian-Russian speech depending on the region and the characteristics of speakers (Hentschel et al., 2014).

Creating a corpus is a promising modern method of studying code-switching, as it allows us to see code-switching in a broader linguistic context and to quantify language use.

Dedicated corpora of code-switching:

- English and Hindi (Dey et al., 2014)
- English and Welsh (Deuchar et al., 2018),
- German and Turkish (Çetinoğlu, 2016),
- Estonian and Russian (Zabrodskaja, 2009), etc.

The experience of compiling code-switching corpora based on **parliamentary texts** already exists:

- Dutch-French speeches from the Belgium Federal texts (Marx, 2010)
- Bilingual Corpus of Basque Parliamentary Transcriptions (Escribano, 2022).

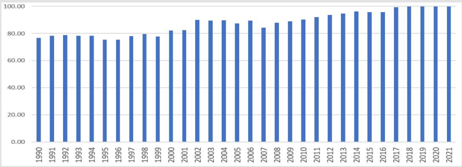
The BasqueParl corpus contains bilingual transcripts in Basque and Spanish and represents the bilingual discourse of the Basque Parliament. It is designed for the automatic analysis of political discourse, including the use of languages and their correlation with entities. BasqueParl shows that there has been no significant change in the amount of bilingualism in parliament over the period 2012-2020 which is covered by the corpus.

Corpus description

The corpus of the Verkhovna Rada (the Ukrainian monocameral 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.

The corpus consists of different sets of text files grouped either by year or by speaker. 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.

A specific feature of the corpus is that it represents a bilingual Ukrainian-Russian discourse with different shares of Ukrainian, Russian, and mixed speech in different years. The Ukrainian language prevails in the corpus, and its share was increasing over the years: from a minimum of 76% in 1995 to 100% in 2018-2020.



Types of code-switching

• Ukrainian speakers insert phraseology or quotations in Russian (henceforth red for the text recorded in Russian)

Dear colleagues, after this discussion, I have an impression that can be characterized by a well-known phrase [from a 19-century Russian play]: "Make noise, friends, make noise!" (Yuriy Solomatin, 2003).

• Russian speakers insert the names of laws and documents in Ukrainian

We are submitting for your consideration a draft law of Ukraine on amendments to certain legislative acts of Ukraine regarding the bankruptcy of mining enterprises (Victor Turmanov, 2003).

• Unmotivated heavy mixing of Russian and Ukrainian (Surzhyk)

By our way, we expect a significant increase in healthcare costs, as I have already mentioned. The total consolidated budget expenditures on healthcare are going to increase by one and a half times. In addition, a number of targeted programs the government envisages, provides to finance, including, by the way, a possible increase in the price of medicines (Mykola Azarov, 2003).

• The language distinguishes between the official position proclaimed in Ukrainian (possibly read from notes) and personal opinions added in Russian;

Dear Members of Parliament! The Government of Ukraine supports the adoption by the Verkhovna Rada of the draft Law of Ukraine on compulsory insurance of civil liability of vehicle owners in the first reading. This is the official position.
But as a representative of two previous convocations, I would like to add that I first presented a similar draft law here myself back in 1996. Since then, our two convocations have spent time in discussions around this project, so to speak, in search of perfection. And I hear now that the same arguments are being put forward again, roughly (Victor Suslov, 2003).

• Triggered code-switching. In the first example, the speaker switches from Russian to Ukrainian after pronouncing the name of an official document in Ukrainian. The second speaker switches from Ukrainian to Russian after using Russian phraseology.

Dear Vladimir Mikhailovich, Gennady Borisovich! I would like to ask you to include in **the list of objects** the city of Kremenchuk and the city of Zolotonosha. These two cities are not on the list, and the problem is very acute in these two cities. Thank you (Vasyl Havryluk, 2003).

And I want to ask again whether the Ministry of Finance has considered the possibility of canceling certain tax privileges that would bring additional budget revenue. But there is a situation where we have the fuel and energy complex, you know, as a **milk cow**, which basically provides certain resources today when we consider increasing revenues, without thinking about the fact that there was a debt for many years, which in fact did not solve any financial issues in the budget in the future. Thank you (Valeriy Konovaluk, 2003).

• Another language marking quoted speech.

They took me into this ring of workers, a lot of people, (...) and here you are, a representative, standing in front of them one on one. And they put pressure on you: you're not doing anything there, you're not giving us money, you're all just gathering there and sitting! And I say: hold on, hold on, what fraction am I in, everyone is in opposition to the President now, and who am I?" (Olga Ginzburg, 2003).

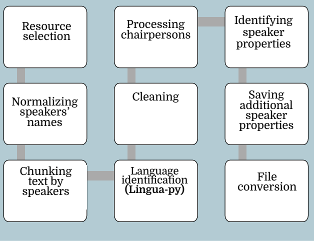
• Switching to another language to illustrate a tolerant attitude to linguistic diversity.

As for the mother tongue, I believe that the mother tongue is the language of the family in which a person was brought up. And in general, let's be tolerant when it comes to both Russian and Ukrainian. We should not politicize this issue (Hennady Vasilyev, Chairman, 2003).

Processing the corpus

When building the corpus we process the speakers and add their party affiliation, annotating the language spoken by the speaker. The sentence was chosen as the unit of annotation. Such modules for determining the language as **CLDv3: Compact Language Detector v3** (Google company), **LangDetect**, **Spacy-langdetect**, **fastText**, **Lingua-py** (ultimately chosen) were tested, and none of these modules showed the desired accuracy. These libraries work poorly with short sentences like "*Djakuju. Levčenko, VO "Svoboda*"" ("Thank you. Levchenko, VO "Svoboda""), "*Vraxovana častkovo*" ("Partially taken into account"), etc., but they do quite well with long sentences.

The transcripts was processed and manually checked for the quality of the split by speakers and language detection.



Resource selection
Processing chairpersons
Identifying speaker properties
Normalizing speakers' names
Cleaning
Saving additional speaker properties
Chunking text by speakers
Language identification **(Lingua-py)**
File conversion

Diachrony

	2007-2012 (6th convocation)	2012-2014 (7th convocation)	2014-2019 (8th convocation)
Ukrainian	67,4%	70,1%	68,1%
Russian	2,9%	2,5%	2,5%
Bilingual	29,7%	27,4%	29,4%

The proportional ratio of Russian-speaking, Ukrainian-speaking, and bilingual speakers in the work of the Parliament of the 6th, 7th, and 8th convocations.

Example of a dialogue

An example of the Ukrainian Verkhovna Rada's transcripts, annotated by language. The transcripts contain dialogues, sometimes in two languages

Speaker's name

Chairperson's identity is indicated once at the start of each day. In the corpus it is annotated in a separate task

```
<lang = "uk">БОНДАР Б.В.</lang>
<lang = "uk">Доповідь закінчив.</lang>
<lang = "uk">Готовий відповісти на пит
<lang = "uk">ГОЛОВУЮЧИЙ.</lang>
<lang = "uk">Дякую, пане генерале.</lang
<lang = "uk">Будь ласка, від фракцій по одному п
<lang = "uk">Будь ласка, Борислав Береза.</lang>
<lang = "uk">Від фракції Радикальної партії передали слово.</lang>
<...>
<lang = "uk">Будь ласка, "Опозиційний блок", Дунаєв Сергій
Володимирович.</lang>
<lang = "uk">ДУНАЄВ С.В.</lang>
<lang = "ru">Прошу передать слово Звягильскому.</lang>
<lang = "uk">ГОЛОВУЮЧИЙ.</lang
<lang = "uk">Юхим Звягільський, будь ласка, народний депутат.</lang>
<lang = "uk">ЗВЯГІЛЬСЬКИЙ Ю.Л.</lang>
<lang = "ru">Уважаемый генерал, а как по
работают все четыре смены?</lang>
<...>
<lang = "ru">Наверное же, надо найти такой режим работы, чтобы шахтеры,
которые работают круглосуточно, могли доставить их и на работу, и с
работы домой.</lang>
<lang = "ru">Как этот вопрос решить?</lang>
<lang = "uk">БОНДАР Б.В.</lang> <lang = "uk">Я хотів би наголосити, що
з цією метою в штабі АТО створено Координаційний центр і в секторах
відповідно на напрямках координаційні групи, які зі складу представників
місцевих органів влади, МВС, СБУ, які здійснюють і забезпечують пропуск
населення для виконання завдань і по роботі, і
<...>
<lang = "uk">(Шум у залі) Прошу.</lang
<lang = "ru">(Оплески)</lang>
```

Question in Russian

Answer in Ukrainian

Indications like 'noise' and 'applause'.
NB! wrong language identification in a short chunk

Conclusion

We traced 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.

The use of both Russian and mixed speech correlates with the membership in Communist or Regions parties.

We tried to analyze whether normative documents (laws, regulations) and the general political situation influence the actual use of languages in the Rada. It turned out the influence of these legal acts was little noticeable.

- **1989** – Law "On Languages in the Ukrainian SSR": in the Ukrainian SSR the language of work, record keeping, and documentation, as well as relations between state, party, public bodies, enterprises, institutions, and organizations is the Ukrainian language.
- **2010** – Regulations of the Verkhovna Rada of Ukraine defined the state language as the working language of the Verkhovna Rada, its bodies, and officials, and speeches in another language were allowed only to foreigners and stateless persons.
- **2012** – Law of Ukraine "On the Principles of State Language Policy" (the so-called Kivalov-Kolesnichenko language law) allowed speaking in the parliament in any language
- **February 2018** – the Kivalov-Kolesnichenko law was repealed.

But 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).

References

Özlem Çetinoğlu. 2016. A Turkish-German Code-Switching Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* Portorož, Slovenia.

Margaret Deuchar, Peredur Davies, Kevin Donnelly. 2018. *Bulding yje Siarad Corpus: Bilingual conversations in Welsh and English,* Amsterdam: Benjamins.

Anik Dey and Pascale Fung. 2014. A Hindi-English Code-Switching Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Nayla Escribano, Jon Ander González, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez-de-Viñaspre, Rodrigo Agerri. 2022. BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions. *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)* Marseille, France. https://aclanthology.org/2022.lrec-1.361.pdf

Gerd Hentschel, Jan Patrick Zeller, Sviatlana Tesch. 2014. Das Oldenburger Korpus zur weißrussisch-russischen gemischten Rede: OK-WRGR, https://uol.de/ok-wrgr

Volodymyr Kulyk. 2022. Language and identity in Ukraine at the end of 2022. (in Ukrainian) 07.01.2023 https://zbruc.eu/node/114247

Maarten Marx and Anne Schuth. 2010. DutchParl. The parliamentary documents in Dutch. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC'10*), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/263_Paper.pdf

Anastasia Zabrodskaja. 2009. Evaluating the Matrix Language Frame model on the basis of a Russian—Estonian code switching corpus. *International Journal of Bilingualism*, 13(3), 357–377.