The Parliamentary Code-Switching Corpus: Bilingualism in the Ukrainian Parliament in the 1990s-2020s

Olha Kanishcheva
University of Jena
kanichshevaolga@gmail.com

Maria Shvedova
University of Jena
National University "Lviv Polytechnic"
mariia.o.shvedova@lpnu.ua

Tetiana Kovalova
University of Jena
V. N. Karazin Kharkiv National University
tvkovalova@karazin.ua

Ruprecht von Waldenfels
University of Jena
ruprecht.waldenfels@uni-jena.de

Abstract

We describe a Ukrainian-Russian code-switching corpus of Ukrainian Parliamentary Session Transcripts. The corpus includes speeches entirely in Ukrainian, Russian, or various types of mixed speech and allows us to see how speakers switch between these languages depending on the communicative situation. The paper describes the process of creating this corpus from the official multilingual transcripts using automatic language detecting and publicly available metadata on the speakers. On this basis, we consider possible reasons for the change in the number of Ukrainian speakers in the parliament and present the most common patterns of bilingual Ukrainian and Russian code-switching in parliamentarians' speeches.

1 Introduction

As a result of Ukraine's long history of political dependence on Russia, first as part of the empire, then throughout its Soviet history, 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, 2023).

In the 20th century, the issue of Ukrainian-Russian bilingualism was the subject of many studies by Ukrainian linguists with a focus on deviations from Russian normative use and interference from Ukrainian. In the 21st century, the focus was mainly on sociolinguistic studies concerning the distribution of both languages and the tasks of supporting the Ukrainian language and pushing back stigmatized mixed Ukrainian-Russian speech known as Surzhyk. The main research methods applied to Ukrainian-Russian bilingualism were interviews and questionnaires, as well as analysis of individual texts, dictionaries, and normative sources.

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 and Reuther, 2020; Palinska and Hentschel, 2022).

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. Dedicated corpora of code-switching have been created for English and Hindi (Dey and Fung, 2014), English and Welsh (Deuchar et al., 2018), German and Turkish (Özlem Çetinoğlu, 2016), Estonian and Russian (Zabrodskaja, 2009), and many more.

This study aims to create a corpus of Ukrainian-Russian code-switching based on transcripts of Ukrainian parliamentary sessions. These transcripts include not only parliamentary speeches, but also discussions between the speakers. This presents a rich bilingual discourse with speakers using Ukrainian, Russian, or different kinds of mixed speech and switching between these languages depending on the communicative situation. This corpus will allow us to improve our understanding of common switching patterns found in Ukrainian parliamentary speeches.

The remainder of this paper is organized as fol-

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 79–90

lows: Section 2 discusses related work. specifically existing code-switching corpora, their features, and research related to these corpora, and presents a selection of the most important recent work in this domain. In Section 3 we present the code-switching corpus of Ukrainian parliamentary session transcripts and go into detail describing the main features of this corpus. The pre-processing, normalization, and processing steps during corpus compilation are given in Section 4. Here, we present the results of the separation of transcript into speakers and language identification. In Section 5, we present an analysis of the transcripts regarding the speaker's language, as it relates to normative documents and political events, showing possible reasons affecting the use of Ukrainian in parliament. Here we also present some typical cases of code-switching. The last Section 6 finalizes the paper and suggests some future works and improvements.

1.1 Research Tasks on Ukrainian Code-Switching Corpus

The transcripts of parliamentary sessions have become the material for numerous linguistic corpora. The CLARIN[1] collection contains 31 such corpora for different languages. (Kryvenko, 2018) reports the creation of a corpus of Ukrainian parliamentary texts for discourse analysis. The corpus does not have language annotation and consists of 1.26 million tokens of different types of parliamentary texts from 2002-2017 (parliamentary news, minutes of plenary sittings, hearings and committees' meetings, Speaker's addresses, committee agendas, reports, announcements, etc.). In 2021, a corpus of about 70 million tokens of Verkhovna Rada plenary session transcripts from 1990-2020 was added to GRAC.v.12[2], with the text in non-Ukrainian languages automatically removed (Starko et al., 2021). The parliamentary transcripts are a ready-to-research record of spoken language, which is of considerable size and available for download from an open source. An important advantage of such texts is also the publicity of information about the speakers, which allows for the most detailed annotation of the corpus and free access to their biographies in the process of deeper linguistic research. Parliamentary corpora can be used not only in the field of linguistic research but also in the social sciences, for various studies of parliamentary discourse. Since the Verkhovna Rada transcripts corpus contains texts in Ukrainian, Russian, and bilingual mixed speech, this corpus can be used to study code-switching, both from a sociolinguistic and psycholinguistic perspective.

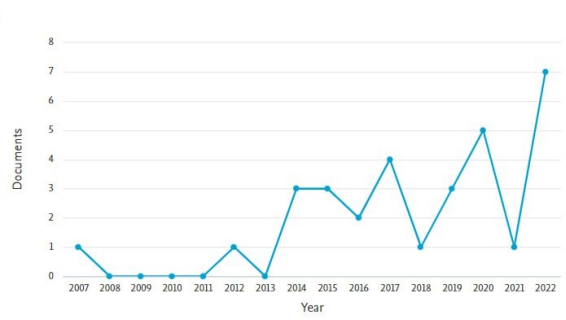

Figure 1: The number of publications related to code-switching corpora in the Scopus database.

2 Related Work

Over the past few years, research in the field of code-switching corpora has increasingly attracted the attention of researchers. Figure 1 shows the number of publications related to code-switching corpora in the Scopus database with overall dynamics increasing every year.

A lot of research is devoted to corpora with audio tracks, and these corpora are used to improve the quality of speech recognition with mixed corpora. Modipa and Davel (2022) present two reference corpora for the analysis of Sepedi-English code-switched speech in the context of automatic speech recognition. Sreeram et al. (2019) describe the collection of a Hinglish (Hindi-English) code-switching database at the Indian Institute of Technology Guwahati (IITG) which is referred to as the IITG-HingCoS corpus. Dau-Cheng et al. (2015) introduce the South East Asia Mandarin-English corpus, a 63-h spontaneous Mandarin-English code-switching transcribed speech corpus suitable for LVCSR and language change detection/identification research. The corpus consists of recordings of unscripted interviews and free conversations by 157 Singaporean and Malaysian speakers who speak a mixture of Mandarin and English switching within sentences.

Some researchers consider code-switching corpora from the point of view of psycholinguistics and investigate the reasons for switching from one language to another. Beatty-Martínez et al. (2020)

[1] https://www.clarin.eu/resource-families/parliamentary-corpora

[2] http://uacorpus.org/

show that code-switching does not always involve additional effort and resources. Deuchar (2020) presents the differentiation of code-switching from borrowing, the methods for evaluating competing models of grammaticality in code-switching, and the importance of studying variables as well as uniform patterns in code-switching.

Thus, researchers use code-switching corpora for various tasks, investigating different aspects from speech recognition to psycholinguistic reasons for switching between languages.

In our work, we study a code-switching corpus of parliamentary speech. Several such code-switching corpora based on parliamentary texts already exist, including the 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). The DutchParl corpus (the Parliamentary Documents in Dutch) contains the Belgian Federal documents, bilingual French-Dutch texts which are presented in the original French or Dutch and contain an aligned translation in the second language. The BasqueParl corpus contains only original 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 [p. 3387].

A 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, gradually reaching 100% Ukrainian in the second half of the 2010s. The language policy in the Ukrainian parliament has been a hot political issue for all these years, and the actual use of languages has varied depending on the political situation. The corpus shows the history of the existence and decline of postcolonial bilingualism in parliamentary discourse.

3 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[3]. The timespan starts even before Ukrainian independence when Verkhovna Rada was an institution of a Soviet republic. The size of the initial data is about 70 million tokens. A specific feature of the corpus is that it represents a bilingual Ukrainian-Russian discourse with different shares of Ukrainian, Russian, and bilingual 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.

The parliamentary speeches and remarks are recorded literally, in the language actually spoken, and language mixing is also accurately reproduced in the transcript. This accuracy allows us to analyze the use of a particular language in dialog and its correlation with other interlocutors' language and the session's topic.

The corpus consists of text files ("txt" format), organized by speaker, that is, text files that contain all the utterances of each member of parliament for the year made both in speeches and in the discussion; this is 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 being annotated by age and the political party of a given speaker, as well as by the administrative region of Ukraine (or by another country if applicable) where they were born and educated.

4 Corpus Preparation

The transcript of parliamentary sessions is a single file that contains all parliamentary sessions for the whole year. An example excerpt from the 2013 transcript is given in Appendix A.

From this fragment, it can be seen that the speakers are written in capital letters with initials. Sometimes this is the name and middle name, i.e. two initials, and sometimes only one initial, only the name is given. At the beginning of each session, the Chairman or *head* is announced who is in charge of the meeting, and further in the text he is referred to merely as "HOLOVUJUČYJ" (Presiding).

Also, the text may contain timestamps (16:11:06), quotations (for example, "...In accordance with Article 13 of the Law of Ukraine "On the Status of People's Deputy of Ukraine..."), and remarks (for example, "Splashes", "Noise in the hall").

[3] https://iportal.rada.gov.ua/meeting/stenogr

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

Table 1: The quantitative data of our corpus.

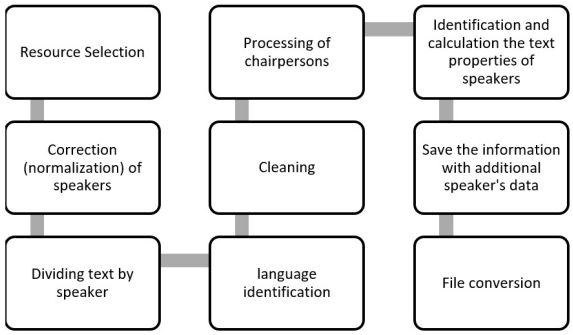

Figure 2: A general pipeline of parliament transcripts processing.

In this work, transcripts from 2010 to 2019 were processed. General information on these files is presented in Table 1.

As a result of the transcript analysis, a general approach to the processing of parliamentary transcripts was formed, which has the stages presented in Figure 2.

Due to the fact that the stenographer allows for variations in the spelling of the names and surnames of speakers, such as a large number of spaces, or spaces between initials, before processing the entire file, we normalize all speakers, i.e. we bring them to the form Surname N.P. This helps to further significantly reduce the number of incorrect files for each incoming.

Then we divide the entire transcript by speakers, namely members of parliament, invited ministers, etc. As a result, a separate file was generated for each speaker. An example of such a file is given in Appendix B.

Speaker separation is done automatically on the basis of the transcription. Sometimes the transcriber makes mistakes in spelling the last name or initials, which results in the software recording several speakers instead of one. Such mistakes have been corrected manually. For example, the surname «Arzhevityn» can be misspelt as «Azhevityn» and without manual verification, it is quite difficult to understand whether this is a real surname or whether it is a mistake. This could be automated if only members of parliament were present in the transcript, but since there are invited people, this cannot be implemented.

Next, we carry out the identification of the head at a meeting in parliament. He/She is mentioned once at the beginning of the meeting. He/She can also change within one convocation. Therefore, the statements of head for one convocation may refer to different people.

We then clean up the speakers' lines of timestamps and remarks.

At the next stage, we determine the language(s) of each speaker on the sentence level. To this end, tried out different modules for determining the language: CLDv3: Compact Language Detector v3 (Google company)[4], LangDetect[5], Spacy-langdetect[6], fastText[7], Lingua-py[8] were tested, and none of these modules showed the desired accuracy. For example, let's consider sentences for which the language was specified as Ukrainian, but they are written in Russian.

- "Davajte slovo" ("Give us the floor ")
- "No ved' tak že nespravedlivo!" ("But it's just as unfair!")
- "Davajte vernem!" ("Let's return!")
- "No ja vam skažu tak." ("But I'll tell you this.")
- "Ja choču poblagodarit' gospodina Grojsmana." ("I want to thank Mr. Hroysman.")

The sentence *"Davajte slovo"* can be both in Russian and in Ukrainian, it is really difficult for the system to determine the language of the sentence based only on the spelling, and in this case, it completely matched. Only phonetic notation can help here. The remaining examples are written in Russian but identified as Ukrainian.

We chose the Lingua-py library, as it made fewer identification errors than the rest of the tested libraries. All experiments to evaluated the accuracy of language detection were carried out manually. However, this library was also making mistakes.

To determine to what extent a speaker uses both languages, we first identified the language used in each sentence using the above-mentioned module and then summed up the number of tokens for

[4] https://github.com/google/cld3
[5] https://pypi.org/project/langdetect/
[6] https://pypi.org/project/spacy-langdetect/
[7] https://fasttext.cc/docs/en/language-identification.html
[8] https://github.com/pemistahl/lingua-py

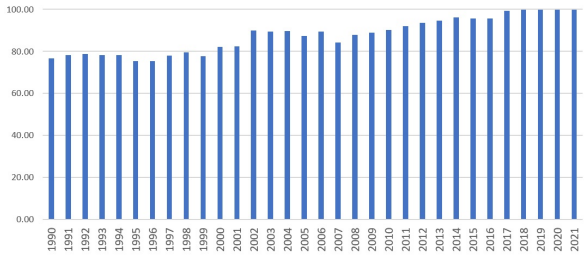

Figure 3: A quantitative report of parliament speeches by language for each year (1990-2021).

each language per speaker. Then the percentage of Ukrainian in the speaker's speech was determined in tokens. Sentences shorter than 5 tokens were not considered, since language recognition modules often make mistakes in short sentences.

Next, we determined which party the representative belongs to, as well as some statistical characteristics of the text, such as the number of tokens offered, etc. In the future, we also plan to add the year and place of birth of the parliament members, in order to check whether the age and region of birth influence language preferences and switching between languages.

5 Data Analysis

5.1 Research Tasks on Ukrainian Code-Switching Corpus

To analyze the use of languages in the Ukrainian parliament, we divided the texts into separate files by the speaker and the year of the three sessions of the parliament. The data are grouped by convocations:

- Verkhovna Rada of Ukraine of the 6th convocation (2007-2012) (II half).
- Verkhovna Rada of Ukraine of the 7th convocation (2012-2014).
- Verkhovna Rada of Ukraine of the 8th convocation (2014-2019).

For each convocation, the number of speakers using Ukrainian, Russian, or both languages was counted.

A quantitative report of parliament speech by language for each year (1990-2021) is given in Figure 3. 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. Thus, the 1989 Law "On Languages in the Ukrainian SSR" determined that 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 (Law, 1989). The Regulations of the Verkhovna Rada of Ukraine adopted in 2010 defined the state language as the working language of the Verkhovna Rada, its bodies, and officials. Speeches in other languages were allowed only to foreigners and stateless persons (Regulations, 2010).

	6 conv.	**7 conv.**	**8 conv.**
Ukrainian	67,4%	70,1%	68,1%
Russian	2,9%	2,5%	2,5%
Bilingual	29,7%	27,4%	29,4%

Table 2: The proportional ratio of Russian-speaking, Ukrainian-speaking, and bilingual speakers in the work of the Parliament of the 6th, 7th, and 8th convocations.

In 2012, the Kivalov-Kolesnichenko law was adopted, which allowed the use of not only Ukrainian but also other working languages in the parliament (Law, 2013). This law caused significant public resonance and was revoked in February 2014 after the Russian invasion. In February 2018, this law was declared unconstitutional.

In 2019, a new Law "On Ensuring the Functioning of the Ukrainian Language as the State Language" was adopted, which established the mandatory use of Ukrainian in the official sphere (Law, 2019).

As can be seen from Figure 3, in 2007, the smallest share of the use of the Ukrainian language by speakers of the parliament was found. The increase in the use of the Russian language in the parliament in 2007 may be related, on the one hand, to the campaign in the Verkhovna Rada against President Viktor Yushchenko, who supported a pro-Western course and derussification, and on the other hand, to Ukraine's ratification of the European Charter for Regional and Minority Languages. After the adoption of the Charter, the so-called "parade of linguistic sovereignty" took place in Ukraine, when a number of local councils of the eastern and southern regions of Ukraine, violating the Constitution of Ukraine and the Law of Ukraine "On Local Self-Government", declared Russian the regional language in their respective territories.

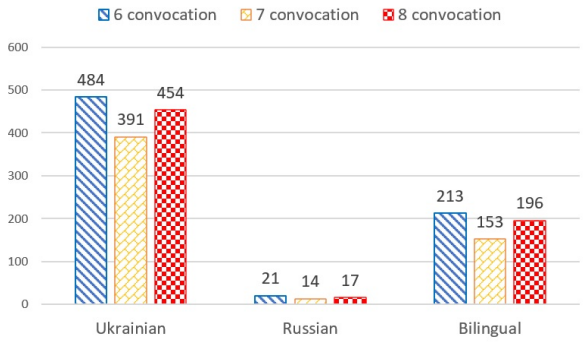

Figure 4: A quantitative report of parliament speakers by language for each convocation (6, 7, and 8 convocations).

Researchers note that the struggle between the pro-Ukrainian and the pro-Russian approach to state language policy intensified after the March 2006 parliamentary elections won by the pro-Russian Party of Regions and gradually entered a heated phase. The efforts of the pro-Ukrainian wing in the executive authorities to support the Ukrainian language encountered opposition from the deputies of the Party of Regions in the parliament and some local councils. In parallel with this, political actions were taking place to give the Russian language the status of the second state language (Marusyk, 2015; Masenko, 2018; Shumlyanskyi, 2007; Skvirska, 2008).

The language distribution by political parties is presented in Appendix C. This appendix provides a table that shows to which extent using a particular language goes along party lines and how often representatives of different parties use both languages.

We assumed that the proportions of Ukrainian speakers, Russian speakers, and switchers in each convocation would depend on the composition of the parties in it; moreover, we hypothesized that the pro-government parties would speak Ukrainian more than others. The largest parties in the parliament of the 6th convocation were the Party of Regions and Yulia Timoshenko's Bloc, of the 7th convocation, Party of Regions and the All-Ukrainian Association "Batkivshchyna", of the 8th convocation, People's Front and Petro Poroshenko's Bloc (Protocol, 2007; Protokol, 2012; Protocol, 2014). A quantitative report of parliament speakers by language for each convocation (6, 7, and 8 convocations) is given in Figure 4, the proportion practically does not change (Table 2).

It should be noted that among the members of the Party of Regions that openly proposed the status of Russian as the second state language the share of language switchers compared to Ukrainian speakers in all convocations was significant. Thus, in the Regions' fraction within the Parliament of the 6th convocation, the number of code-switchers is higher than the number of Ukrainian-speakers (77 and 63 respectively). In the 7th convocation the respective numbers are close: 50 and 53.

The absolute leader among language switchers in the work of the 6th and 7th convocations of the Parliament, as can be seen from the data presented in the appendices, is the Communist Party of Ukraine, known for its pro-Russian political platform (in the 8th convocation, the party was actually legally banned).

The rather large share of code-switchers in comparison with Ukrainian speakers in the Petro Poroshenko Bloc of the 8th convocation is noteworthy. Most likely, this is due to the fact that the head of the bloc and its members advocated for the regional status of the Russian language, distinguishing it from other national minority languages. See, for example, draft law No. 4178a of 26 June 2014, where Petro Poroshenko proposed to amend Article 143 of the Constitution of Ukraine, which would allow local authorities to change the status of a language with a special emphasis on Russian (DraftLaw, 2014). Petro Poroshenko later changed his position on the state language as reflected in his speech at the Verkhovna Rada on 20 September 2018, where he expresses strong support for the Ukrainian language. This corresponds to the achievement of monolingualism in the Ukrainian parliamentary discourse in general, as shown in Figure 3.

5.2 Cases of Code-Switching

In this section, we describe the most common types of combined use of Ukrainian and Russian we found in parliamentarians' speeches; note that at this point, we adduce data only from 2003, which has been manually annotated. Below, each type is illustrated with examples from the transcripts. Ukrainian text is given in standard font, and Russian text is in cursive font. All examples have been translated into English.

- Ukrainian speakers insert phraseology or quotations in Russian.

Šanovni kolehy, u mene pislja toho jak u nas vidbuvajet'sja ce obhovorennja, skladajet'sja take

vraženna, jake možna oxarakteryzuvaty vidomym vyslovom: *"Šumim, bratcy, šumim!"* (Jurij Solomatin, 2003)[9].

Dear colleagues, after this discussion, I have an impression that can be characterized by a well-known phrase: *"We make noise, friends, we make noise!"* (Yuriy Solomatin, 2003. The phrase is based on a quotation from *Woe from Wit*, the classical 19th-century Russian play by A. Griboyedov. Here and below we italicize the text that appears in the original Russian).

- Russian speakers insert the names of laws and documents in Ukrainian.

Na vaše rassmotrenie vnositsja proekt zakona Ukrainy pro vnesennja zmin do dejakyx zakonodavčyx aktiv Ukraïny ščodo bankrutstva hirnyčnyx pidpryjemstv (Viktor Turmanov, 2003).

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).

- Russian speakers insert Ukrainian legalese technical terms and clichés.

Predlagaju v cilomu. *Mnogo golosov "za". Vse* zauvaženna *učteny* (2003).

I propose [to adopt the draft law] as a whole. *Many votes in favor. All* the comments *have been taken into account* (2003).

- Ukrainian speakers insert words in Russian.

Ce ne prjamyj zv'jazok, a ce poserednij... *kosvennyj*... poserednij zv'jazok (Mykola Poliščuk, 2003).

This is not a direct connection, but an indirect... *intermediate*... indirect connection (Mykola Polishchuk, 2003).

- Unmotivated heavy mixing of Russian and Ukrainian.

My, do reči, peredbačajemo značne zbil'šennja vytrat na medycynu, *ja ob ėtom uže govoril. Sovokupnye raschody konsolidirovannogo bjudžeta na medicinu vozrastajut u nas v poltora raza*. Krim toho, cilyj rjad cil'ovyx prohram *pravitel'stvo* peredbačaje, *predusmatrivaet na finansirovanie, v tom čisle, kstati, i na* možlyve pidvyščennja likars'kyx zasobiv v cini (Mykola Azarov, 2003).

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 speaker switches languages for stylistic purposes, distinguishing between the official position proclaimed in Ukrainian and personal opinions added in Russian.

Šanovnye *narodnye deputaty*! Urjad Ukraïny pidtrymuje sxvalennja Verxovnoju Radoju proektu Zakonu Ukraïny pro obov'jazkove straxuvannja cyvil'no-pravovoï vidpovidal'nosti vlasnykiv transportnyx zasobiv v peršomu čytanni. Ce oficijna pozycija.

No kak narodnyj deputat dvuch predyduščich sozyvov, ja choču dobavit', čto vpervye podobnyj proekt ja sam dokladyval zdes' ešče v 1996 godu. S tech por dva našich sozyva proveli vremja v diskussijach vokrug ėtogo proekta, tak skazat', v poiskach soveršenstva. I ja sejčas slyšu, čto vidvigajutsja vnov' te že samye argumenty, primerno (Viktor Suslov, 2003).

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.

Uvažaemyj Vladimir Michajlovič, Gennadij Borisovič! Ja chotel by prosit' vključit' do **pereliku ob'jektiv** šče misto *Kremenčug* ta misto

[9]The text is presented in transliterated form according to the original transcript, without typographical or other corrections.

Zolotonošu. Cyx dvox mist nemaje v pereliku, a problema duže hostra v cyx dvux mistax je. Djakuju (Vasyl' Havryljuk, 2003).

Dear Vladimir Mikhailovich, Gennady Borisovich! I would like to ask you to include in the **list of objects** the city of *Kremenchug* 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).

I ja xoču šče raz spytaty, čy rozhljadalasja možlyvist' Ministerstvom finansiv skasuvaty okremi podatkovi pil'hy, jaki b daly dodatkovi doxody bjudžetu. Ale vynykaje taka sytuacija, ščo u nas palyvno-enerhetyčnyj kompleks, znajete, jak ***dojnaja korova**, kotoraja v principe i obespečivaet segodnja opredelennye resursy, kogda my rassmatrivaem uveličenie dochodom, ne dumaja o tom, čto byla mnogie gody ta nedoimka, kotoraja po suti dela absoljutno ne rešala absoljutno nikakich finansovych voprosov i v bjudžete v dal'nejšem. Spasibo* (Valerij Konovaljuk, 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).

- Code-switching in a dialog under the influence of the interlocutor's speech.

HOLOVA. (...) Propozycija komitetu jaka? Bud' laska, Vasyl' Petrovyč.

CUŠKO V. P. [mostly Russian-speaking]. Propozicija *komiteta – podderžat' v pervom čtenii* (2003).

CHAIR. (...) What is the Committee's proposal? Please, Vasyl Petrovych.

TSUSHKO V. P. The Committee's proposal *is to support it in the first reading* (2003).

- Language switching is used to mark quoted speech.

Vony vzjaly mene u take kil'ce - ce robitnyky, masa ljudej, (...) a ty, deputat, stoïš pered nymy odyn na odyn. I vony na tebe tysnut': *čto ty tam ničego ne delaeš', den'gi nam ne dajut, vy tam vse sobiraetes' i sidite!* A ja kažu: ty pidoždy, ty pidoždy, ja v jakij frakciï naxodžusja, vsi zaraz v opozyciï do Prezydenta, a xto ja taka? (Ol'ha Hinzburh, 2003).

They encircled me, they are workers, a lot of people, (...) and here you are, a representative, standing in front of them face to face. And they put pressure on you: *why don't you do anything there, they don't give us any money, you all just gather and sit there!* And I say: hold on, hold on, what deputy group am I in, everyone is in opposition to the President now, and who am I?" (Olha Hinzburh, 2003).

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

A ščo stosujet'sja ridnoï movy, ja tak vvažaju, ščo ridna mova – ce mova rodyny, v jakij vyxovuvalasja ljudyna. *I voobšče davajte tolerantno otnositsja, čto kasaetsja i russkogo i ukrainskogo jazyka. Ne sleduet politizirovat' ėtot vopros* (Henadij Vasyl'jev, holovujučyj, 2003).

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).

The examples presented were taken from the 2003 transcript not yet included in the corpus, however, we think the same types are also to be found in the data from 2010 to 2019. Still, the identification of new types of language switching requires a more detailed analysis and is planned to be carried out in future studies.

6 Conclusions and Future Plans

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.

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. We tried to analyze whether laws and the general political situation affect the actual use of languages in the Council. It turned out that laws are a deterrent to increasing the use of the Russian language in

parliament. In some cases, the influence of political trends on the use of languages can be assumed (for example, 2007, when the increase in the share of the Russian language coincided with the pro-Russian campaign in Ukraine), but this requires 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.

Besides, in the future, it is planned to connect the interface and change the corpus storage format in order to store dialog information and all the necessary metadata.

Limitations

We see the following main limitations at this point in time:

- The error rate in distinguishing Russian and Ukrainian and its impact is not known.
- Due to variations in the input data, the automatic speaker identification needs extensive manual post-editing.
- Different types of code-switching are extremely hard to automatically distinguish.
- In our data collection approach, we combine all utterances of each speaker in a single file. Right now, we therefore cannot automatically distinguish between speakers who use both Russian or Ukrainian alternatively, without mixing within a unit of discourse (bilingual speakers) and speakers who mix languages (code-switching speakers). This will be addressed in later work.
- Right now, we do not take the amount of data of speakers into account. Naturally, speakers with a lot of data are more probable to have text in both languages; this is disregarded right now and its impact is unclear.

Ethics Statement

Our scientific work complies with the ACL Ethics Policy[10]. The corpus was created on the basis of publicly available data.

Acknowledgements

We would like to thank Reviewers for taking the time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions, which helped us to improve the quality of the manuscript.

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

Anne L. Beatty-Martínez, Christian A. Navarro-Torres, and Paola E. Dussias. 2020. Codeswitching: A Bilingual Toolkit for Opportunistic Speech Planning. *Frontiers in Psychology*, 11.

Lyu Dau-Cheng, Tan Tien-Ping, Chng Eng-Siong, and Li Haizhou. 2015. Mandarin-English code-switching speech corpus in South-East Asia: SEAME. *Language resources and evaluation*, 49(3):581–600.

Margaret Deuchar. 2020. Code-Switching in Linguistics: A Position Paper. *Languages*, 5(2):22.

Margaret Deuchar, Peredur Davies, and Kevin Donnelly. 2018. *Building and Using the Siarad Corpus: Bilingual conversations in Welsh and English (Studies in Corpus Linguistics (SCL), 81, Band 81)*. John Benjamins Publishing Co.

Anik Dey and Pascale Fung. 2014. A Hindi-English Code-Switching Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2410–2413, Reykjavik, Iceland. European Language Resources Association (ELRA).

DraftLaw. 2014. Draft Law on Amendments to the Constitution of Ukraine (regarding the powers of state authorities and local self-government bodies) of 26.06.2014 No. 4178a.

[10] https://www.aclweb.org/portal/content/acl-code-ethics

Nayla Escribano, Jon Ander González, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez de Viñaspre, and Rodrigo Agerri. 2022. BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, page 3382–3390, Marseille, France. European Language Resources Association (ELRA).

Gerd Hentschel and Tilmann Reuther. 2020. Ukrainisch-russisches und russisch-ukrainisches Code-Mixing. Untersuchungen in drei Regionen im Süden der Ukraine. *Colloquium: New Philologies*, 5:105–132.

Anna Kryvenko. 2018. Constructing a Narrative of European Integration in the Verkhovna Rada of Ukraine: A Corpus-Based Discourse Analysis. *Cognition, communication, discourse*, 17:56–74.

Volodymyr Kulyk. 2023. Language and identity in Ukraine at the end of 2022. *Zbruch*.

Law. 1989. On languages in the Ukrainian SSR: Law of the Ukrainian Soviet Socialist Republic of October 28, N 8312-11.

Law. 2013. On the principles of state language policy: Law of Ukraine dated july 3, 2012 No. 5029-vi. *Information of the Verkhovna Rada of Ukraine*, No. 23. Art. 218.

Law. 2019. On ensuring the functioning of the Ukrainian language as a state language: Law of Ukraine dated April 25, 2019 No. 2704-viii. *Information of the Verkhovna Rada of Ukraine*, No. 21, Article 81.

Taras Marusyk. 2015. State language policy in Ukraine in the last decade. *Universe*, 3-4:257–258.

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.

Larisa Masenko. 2018. Language conflict in Ukraine: ways of solution. *Ukrainian Language*, 2:20–35.

Thipe I. Modipa and Marelie H. Davel. 2022. Two sepedi-english code-switched speech corpora. *Lang Resources & Evaluation*, 56:703–727.

Olesya Palinska and Gerd Hentschel. 2022. Regional'ny'e osobennosti ispol'zovaniya ukrainsko-russkoj smeshannoj rechi (surzhika) i vliyanie dialektov: pristavki i predlogi VID / OT. *LingVaria*, 34(2):229–253.

Protocol. 2007. Protocol of the Central Election Commission "On the results of the elections of people's representatives of Ukraine".

Protocol. 2014. Protocol of the Central Election Commission "On the results of the elections of the national deputy of Ukraine in the general state multimandate electoral district".

Protokol. 2012. Protocol of the Central Election Commission "On the results of the elections of the national deputy of Ukraine in the general state multimandate electoral district".

Regulations. 2010. About the Regulations of the Verkhovna Rada of Ukraine. *Information of the Verkhovna Rada of Ukraine*, 14-15, 16-17.

Stanislav Shumlyanskyi. 2007. Bilingual threats and chances of bilingualism. *Criticism*, 1-2:5–7.

Vira Skvirska. 2008. "Language is a weapon of politics", or about language problems in post-Soviet Odesa. page 167–195.

Ganji Sreeram, Dhawan Kunal, and Sinha Rohit. 2019. Iitg-HingCoS corpus: A Hinglish code-switching database for automatic speech recognition. *Speech communication*, 110:76–89.

Vasyl Starko, Andriy Rysin, and Maria Shvedova. 2021. Ukrainian Text Preprocessing in GRAC. In *Proceedings of the 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 2, pages 101–104, Lviv, Ukraine.

Anastassia Zabrodskaja. 2009. Evaluating the Matrix Language Frame model on the basis of a Russian-Estonian codeswitching corpus. *International Journal of Bilingualism*, 13(3):357–377.

Özlem Çetinoğlu. 2016. A Turkish-German Code-Switching Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 4215–4220, Portorož, Slovenia. European Language Resources Association (ELRA).

A Sample transcript of a parliamentary session

HOLOVUJUČYJ. Prošu. Narodnyj deputat Klyčko.

16:11:06

KLYČKO V.V.

Dobryj den', dorohi kolehy, xoču šče raz zvernuty uvahu, ščo Verxovna Rada, na žal', prodovžuje ne vykonuvaty svoix funkcij i ne pracjuje jak vona povynna pracjuvaty.

Holova Verxovnoi Rady neodnorazovo nahološuvav na tomu, ščo povynno buty personal'ne holosuvannja. Šče raz xoču nahadaty, ščob my znovu ne povertalysja do praktyky holosuvannja, tak zvane fortepiano čy pianino, koly deputaty bihajut' ta holosujut' za ljudej, jaki ne prysutni v zali. Ce perše. Po-druhe, ja vpevnenyj, ščo

s'ohodnišnij den' my povynni proholosuvaty zakon pro spivpracju Ukraïny ta… jevropejs'koï intehraciï i obov'jazkovo rozhljanuty zakonoproekt pro vybory, miscevi vybory…

HOLOVUJUČYJ. Prošu, dajte zakinčyty.

(Translation:

CHAIR. Please. People's deputy Klitschko.

16:11:06

KLYCHKO V.V.

Good afternoon, dear colleagues, I would like to point out once again that the Verkhovna Rada, unfortunately, continues to not perform its functions and does not work as it should.

The Chairman of the Verkhovna Rada repeatedly emphasized that there should be personal voting. I want to remind once again that we should not return to the so-called piano voting practice, when deputies run and vote for people who are not present in the hall. This is the first.

Secondly, I am sure that today we must vote on the law on cooperation of Ukraine and... European integration and must consider the draft law on elections, local elections...

CHAIR. Please let me finish.

B The example of a file for one parliament speaker

<lang = "uk">Šanovni kolehy, s'ohodni važlyvyj den' dlja Ukraïny.</lang>

<lang = "uk">S'ohodni v Mins'ku vidbudet'sja zasidannja tr'oxstoronn'oï kontaktnoï hrupy.</lang>

<lang = "uk">I my očikujemo vid neï serjoznyx rezul'tativ, može, navit' i proryvu v spravi vrehuljuvannja konfliktu na Donbasi.</lang>

<lang = "uk">Takož s'ohodni v Jevropejs'komu parlamenti vidbudet'sja special'ni sluxannja ščodo vykonannja Mins'kyx domovlenostej.</lang>

<lang = "uk">Cóho domohlasja Ukraïna.</lang>

<lang = "uk">I dlja nas je duže važlyvoju reakcija jevropejs'koï spil'noty na ti hrubi porušennja Mins'kyx domovlenostej, na jaki jdut' rosijs'koterorystyčni vijs'ka na Donbasi.</lang>

<lang = "uk">S'ohodni Mins'ki domovlenosti – ce jedynyj zapobižnyk vid velykoï vijny na Donbasi.</lang>

<lang = "uk">I tomu tak važlyvo nam vsima zasobamy pidtrymaty ïxnje vykonannj.</lang>

<lang = "uk">Takož xoču zvernuty vašu uvahu na te, ščo ljudy na Donbasi vže vtomleni vid toho, ščo tam vidbuvajet'sja.</lang>

<lang = "uk">Včorašni podiï, koly 500 ljudej vyjšly i pišly neozbrojenymy na bandytiv Zaxarčenka z avtomatamy, i skazaly ïm, ščo treba zupynyty te, ščo vidbuvajet'sja tam, zabraty harmaty z ïxnij dvoriv, prypynyty vbyvaty ljudej, prypynyty vijnu, – ce peršyj pryznak toho, ščo vidbuvajet'sja protverezinnja vsjudy, v tomu čysli i na Donbasi.</lang>

<lang = "uk">I my spodivajemosja, duže skoro ukraïnci pokažut' vsim cym najmancjam i bandytam na dveri.</lang>

<lang = "uk">I ešče.[11]</lang>

<lang = "ru">Kak odessit choču obratit' vaše vnimanie na očen' važnyj moment.</lang>

<lang = "ru">Segodnja u nas planiruetsja privatizacija, v tom čisle obsuždaetsja privatizacija "Odesskogo priportovogo zavoda".</lang>

<lang = "ru">Bezuslovno, podderživaja neobchodimost' poiska ėffektivnych sobstvennikov dlja gosudarstvennogo imuščestva, choču obratit' vnimanie, čto my ne možem narušat' zakon.</lang>

<lang = "ru">A u nas est' Zakon "Ob ėkologičeskom audite", kotoryj predupreždaet, čto ljubye dejstvija s takim krajne opasnym predprijatiem kak "Odesskij priportovyj zavod", v sostav kotorogo vchodit krupnejšee v Evrope ammiakochranilišče emkost'ju 120 tysjač tonn.</lang>

<lang = "ru">Vdumajtes' v ėtu cifru – 120 tysjač tonn ammiaka – ne mogut byt' sdelany bez objazatel'nogo ėkologičeskogo audita, k sožaleniju on do sich por ne vypolnen, a v tože vremja predstaviteli pravitel'stva dokladyvajut o planach privatizacii OPZ.</lang>

<lang = "ru">Choču obratit' vnimanie Kabineta Ministrov na neobchodimost' neukosnitel'nogo vypolnenija zakonodatel'stva Ukrainy v sfere ėkologii dlja togo čtoby obespečit' bezopasnost' žitelej Odessy millionnoj i gorodov vokrug nee.</lang>

<lang = "ru">Ved' Odesskij priportovoj zavod nachodit'sja vsego liš' v 15 kilometrach ot pervych mnogokvartirnych domov Odessy i bezopasnost' na nem ėto zalog žizni i zdorov'ja bolee milliona čelovek.</lang>

<lang = "ru">Spasibo.</lang>

C Language picture by political parties in parliament

[11] A typical example of incorrect automatic language detection of a short sentence.

Party	Convocation								
	6 (2007-2012)			7 (2012-2014)			8 (2014-2019)		
	UKR	RUS	Bilingual	UKR	RUS	Bilingual	UKR	RUS	Bilingual
Our Ukraine–People's Self-Defense Bloc / Блок «Наша Україна — Народна самооборона»	38	0	21						
Lytvyn Bloc / Блок Литвина	11	1	4						
Petro Poroshenko Bloc / Блок Петра Порошенка							74	1	46
Yulia Tymoshenko Bloc / Блок Юлії Тимошенко	66	2	40						
All-Ukrainian Agrarian Association "Spade" / ВАО «ЗАСТУП»							1	0	0
All-Ukrainian Union "Fatherland" / ВО «Батьківщина»				62	0	23	2	0	3
All-Ukrainian Union "Freedom" / ВО «Свобода»				24	0	9	2	0	3
Communist Party of Ukraine / Комуністична партія України	2	2	20	12	0	16			
People's Party / Народна партія				0	0	1			
People's Front / Народний фронт							53	0	22
Our Land / Наш край							1	0	0
Self Reliance / Об'єднання «Самопоміч»							13	0	15
Opposition Bloc / Опозиційний блок							4	4	12
Ukrainian Democratic Alliance for Reform of Vitali Klitschko / Партія «УДАР» Віталія Кличка				25	0	8			
Party of Regions / Партія регіонів	63	9	77	53	7	50			
Right Sector / Правий сектор							1	0	0
Radical Party of Oleh Liashko / Радикальна партія Олега Ляшка				0	0	1	10	0	12
Self-nomination				20	1	8	33	4	32
Union / Союз				0	0	1			
Ukrainian Association of Patriots / Українське об'єднання патріотів — УКРОП							3	0	0