Quantifying Code-Switching in a Ukrainian Parliamentary Dataset 1990-2021

Olha Kanishcheva[1,2] Maria Shvedova[3,4]
[1]Heidelberg University, [2]SET University,
[3]National Technical University "Kharkiv Polytechnic Institute", [4]Friedrich Schiller University Jena
Grabengasse 1, 69117 Heidelberg; Mykoly Shpaka St. 3, 03113, Kyiv;
Kyrpychova str. 2, 61002, Kharkiv; Fürstengraben 1 07743, Jena
kanichshevaolga@gmail.com, o.kanishcheva@setuniversity.edu.ua, mariia.shvedova@khpi.edu.ua

Abstract

Analyzing code-switching – the practice of mixing multiple languages in one discourse – remains a significant task in natural language processing (NLP). This study examines the Ukrainian-Russian bilingual context, focusing on quantifying language alternation in a multilingual dataset. We introduce metrics to assess linguistic boundaries and patterns, specifically addressing the complexities of processing texts where Ukrainian and Russian are used interchangeably, including word-level hybridization. Using a corpus of approximately 200,000 tokens derived from parliamentary transcripts (1990-2021), we apply code-switching metrics to identify frequency and patterns of language use. Our findings provide insights into bilingual communication dynamics and can be used to improve language identification models for mixed-language data.

Keywords: code-switching, code-mixing, dataset, Ukrainian, Russian, code-mixing metrics

1. Introduction

Code-switching or code-mixing, the alternating use of multiple languages within discourse, is a widespread phenomenon in multilingual communities. Understanding the patterns and triggers of code-switching is crucial for fields ranging from linguistics and sociolinguistics to NLP (Winata et al., 2023; Doğruöz et al., 2021), where the goal is to model and understand the use of human language. The terms *code-switching* and *code-mixing* are used in studies of multilingual discourse, with distinctions based either on structural differences (seamless mixing vs. distinct switching) or speaker intentionality (intentional switching vs. unintentional mixing) (Hakimov, 2021), but in this paper, we use *code-switching* as an umbrella term encompassing both phenomena, focusing specifically on intra-sentential instances as our dataset consists of isolated sentences.

The structural properties of code-switching have been extensively studied since the 1980s. Poplack (1980) demonstrated that code-switching is not random, but is governed by grammatical constraints, most notably the *Equivalence Constraint*, which predicts that switches occur at points where the surface syntax of both languages is congruent. The *Matrix Language Frame* model (Myers-Scotton, 1993) further argued that one language serves as the grammatical matrix, supplying morphosyntactic structure, while the other contributes lexical insertions – a distinction that proves particularly relevant for morphological mixing. Muysken (2000) proposed a typological framework distinguishing *insertion*, *alternation*, and *congruent lexicalization* as the three fundamental strategies of code-mixing. The latter strategy, in which two languages share grammatical structure that can be filled lexically by either language, is especially pertinent to typologically close language pairs such as Ukrainian and Russian.

The Ukrainian-Russian bilingual community presents a valuable case study for code-switching analysis. As two closely related Slavic languages, Ukrainian and Russian share significant lexical, syntactic, and morphological similarities, making code-switching between them particularly fluid and frequent. Additionally, both languages use the Cyrillic alphabet, which further contributes to the complexity of distinguishing between them, as homographs – words that are spelled identically but differ in pronunciation and sometimes meaning – are common.

In this paper, we aim to fill this gap by presenting a comprehensive analysis of code-switching in a Ukrainian-Russian bilingual dataset from an NLP perspective. We focus on quantifying code-switching through the use of various metrics and conduct an in-depth analysis to uncover patterns at both the lexical and syntactic levels. By leveraging tools from natural language processing, such as symbol n-gram analysis, part-of-speech tagging, etc., we seek to identify the linguistic and contextual factors that influence when and how speakers switch between Ukrainian and Russian.

Our contributions include the calculation of different metrics for quantifying code-switching in bilingual text and the application of these metrics to a dataset comprising spoken Ukrainian-Russian

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 2–12

bilingual data.

The structure of our article is as follows: In Section 2, we provide a detailed description of the dataset, including the source, selection process, and annotation methodology. Section 3 outlines the metrics employed for evaluating code-switching, along with the results obtained from these evaluations. Section 4 presents an analysis of key linguistic features at code-switching points, such as n-grams and parts of speech. Section 5 analyzes collocations that occur at language boundaries and provides detailed examples of code-switching instances. Finally, Section 6 offers the conclusions drawn from our study.

2. Data Description and Statistics

In this study, we compiled a dataset of sentences from Ukrainian parliamentary session transcripts that exhibit code-switching between Ukrainian and Russian. Parliamentary transcripts provide a large volume of contemporary texts published in the public domain and thus often serve as the basis for corpus linguistic research (Erjavec et al., 2024). The transcripts of the Ukrainian parliament's sessions published on the official website[1] also have additional value as linguistic material, as the texts are transcribed verbatim, preserving colloquial syntax, language errors, language switching, etc.

Different transcription nuances in different years need to be taken into account when working with the dataset. Most of the texts were transcribed manually and the work was done quickly, so the texts contain typos, particularly in sentences with code-switching, where the transcriber did not always manage to switch the keyboard layout in time. In the 1990s, transcripts were at least partially edited (in particular, vocative forms were normalized (Shvedova and Lukashevskyi, 2025)). Since 2023, we have noticed signs of automatic transcription, which reduces the value of the material for linguistic research, as the program normalizes the text (up to replacing colloquial words with literary synonyms); this was discovered by comparing transcriptions from recent years with audio recordings.

The primary language of Ukrainian parliamentary transcripts is standard Ukrainian, with a minor presence of Russian that declined annually, with substantial Russian fragments becoming rare after 2017 (Kanishcheva et al., 2023). Nevertheless, code-switching instances, including brief lexical insertions (1-2 words), occur throughout the corpus and are more frequent in earlier transcripts, providing material for creating a bilingual dataset.

To obtain code-switching content, we excluded sentences that were entirely or predominantly in Russian using CleanText.groovy[2]. The remaining sentences were lemmatized with the dictionary-based TagText parser[3] and filtered to retain only those containing more than two out-of-vocabulary words, which typically indicate a mixture of Ukrainian and Russian. This approach yielded a dataset of approximately 150,000 tokens of identified code-switching content. To ensure a balanced distribution for the language identification task, we supplemented this data with an additional 50,000 tokens from previously excluded Russian sentences.

Figure 1 presents the annual distribution of the collected data alongside the frequency of language transitions from 1990 to 2021. The blue bars represent the text frequency (the volume of unique sentences containing code-switching events), while the orange bars illustrate the code-switching frequency (the total count of language switches identified). A higher count of switches relative to the number of sentences, as observed in peak years like 2019, indicates more frequent transitions between Ukrainian and Russian within individual utterances, suggesting a higher density of mixing in those specific periods.

We processed the sentences by tokenizing them and manually labeling the language of each token. The tokens were categorized into five distinct classes: Ukrainian (UK), Russian (RU), Ukrainian-Russian hybridized words (MIX), Others (OTH), Numbers (NUM), and Punctuation (PUNCT). The resulting corpus comprises over 200,000 tokens (Table 1).

While the full annotated corpus comprises over 200,000 tokens to support the classification task, all subsequent statistical calculations and linguistic analyses in this study are conducted on the core code-switching subset of 150,000 tokens.

The dataset was annotated by three native bilingual speakers of Ukrainian and Russian. One annotator (a graduate student) performed the initial annotation with access to expert consultation for difficult cases. Upon completion of this first pass, a systematic review revealed that the task was more nuanced than anticipated. Consequently, we implemented a validation stage in which two expert annotators collaboratively reviewed a substantial portion of the dataset, identified problematic cases, and developed explicit annotation guidelines. Challenging annotation cases are analyzed in detail in (Kanishcheva et al., 2026), such as orthographically identical words at code-switching boundaries, hybrid and morphologically adapted forms, syntactic calques, and the distinction between borrowings, colloquial variants, and dialectal forms.

[1] https://www.rada.gov.ua/documents/Stenbul_pz

[2] https://github.com/brown-uk/nlp_uk/blob/master/doc/README_other.md

[3] https://github.com/brown-uk/nlp_uk

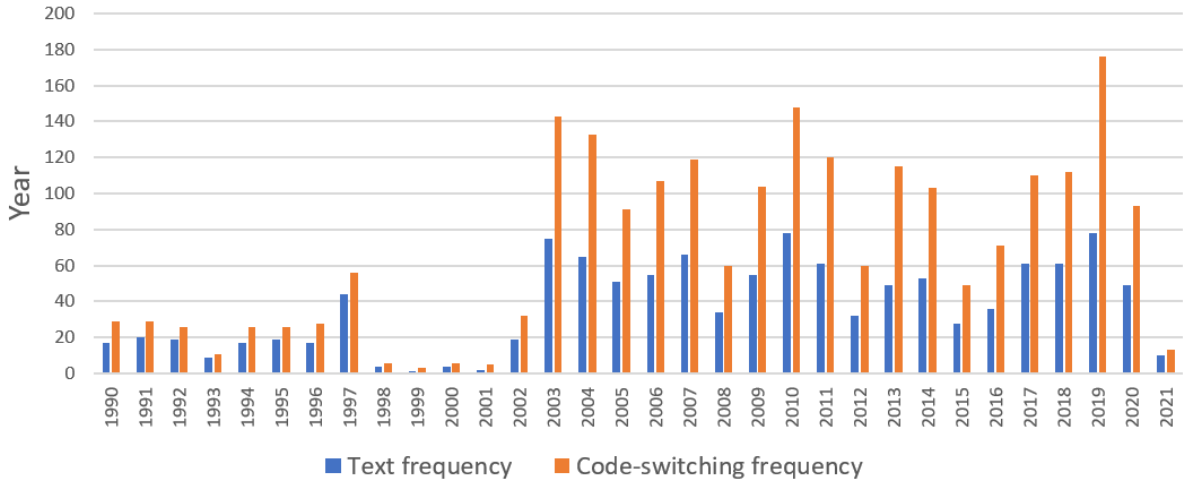

Figure 1: Quantity of sentences containing code-switching instances per year.

The dataset has been published on Zenodo[4] under the Creative Commons Attribution 4.0 International license.

3. Code-Switching Metrics

Code-switching metrics are quantitative measures used to analyze and evaluate patterns of code-switching in bilingual or multilingual discourse (Mave et al., 2018; Guzmán et al., 2017). These metrics provide insights into the frequency, distribution, and structure of code-switching phenomena. Researchers employ various metrics to assess aspects such as the number of code-switching instances, the types of switches (e.g., intra-sentential or inter-sentential), the linguistic levels involved (e.g., lexical, syntactic), and the languages or varieties being switched between. Additionally, code-switching metrics may include measures of proficiency, fluency, and sociolinguistic factors to capture the complexity of code-switching behavior accurately. Overall, the use of code-switching metrics facilitates systematic analysis and comparison of language mixing phenomena across different contexts and populations.

When estimating a code switch in a dataset, the following metrics are usually used:

The **Multilingual Index (M-index)**, developed by Barnett et al. (Barnett et al., 2000) from the Gini coefficient, is a word-count-based measure that quantifies the inequality of the distribution of language tags in a corpus of at least two languages. The M-index is calculated as follows, where $k > 1$ is the total number of languages represented in the corpus, p_j is the total number of words in the language j over the total number of words in the corpus, and j ranges over the languages present in the corpus:

$$M - index = \frac{1 - \sum_{j=1}^{k} p_j^2}{(k-1)\sum_{j=1}^{k} p_j^2}. \quad (1)$$

The index is bounded between 0 (monolingual corpus) and 1 (each language in the corpus is represented by an equal number of tokens).

The **Integration Index** is the approximate probability that any given token in the corpus is a switch point (Guzmán et al., 2017; Guzman et al., 2016). Given a corpus composed of tokens tagged by language $\{l_j\}$, where i ranges from 1 to $n-1$, the corpus size. The I-index is computed as follows:

$$I - index = \frac{1}{n-1} \sum_{1 \le i = j-1 \le n-1} S(l_i, l_j), \quad (2)$$

where $S(l_i, l_j) = 1$ if $l_i \neq l_j$ and 0 otherwise. For a corpus with n tokens, there are $n-1$ possible switch points. It quantifies the frequency of code-switching in a corpus.

The **Code-Mixing Index** is calculated at the utterance level, by finding the most frequent language in the utterance and then counting the frequency of the words belonging to all other languages present in the dataset as illustrated in equation (3) (Das and Gambäck, 2014; Gambäck and Das, 2016).

$$CMI = \frac{\sum_{i=1}^{n}(w_i) - \max\{w_i\}}{n - u}, \quad (3)$$

where $\sum_{i=1}^{n}(w_i)$ is the sum over N languages in the utterance, $\max\{w_i\}$ the highest number of words present from any language, N the number of languages in the utterance, n the number of tokens, and u the number of language-independent tokens. The range of CMI value is [0, 100). If an

[4] https://zenodo.org/records/14724542

Labels	Description	Count	%
UK	Ukrainian words	91,592	41.85
RU	Russian words	82,375	37.65
MIX	Ukrainian-Russian hybridized words	652	0.29
NUM	Numbers	2,123	0.97
OTH	Words in other languages	234	0.1
PUNCT	Punctuation	41,832	19.11

Table 1: Dataset statistics for the language pair Ukr-Rus.

utterance has language-independent tokens or only monolingual tokens, then the corresponding CMI value is 0. A higher value of CMI indicates a higher level of mixing between the languages. CMI-all is an average over all utterances in the corpus and CMI-mixed is an average over only code-switched instances.

Language Entropy (LE): An information-theoretic alternative to the M-index. Measures the number of bits required to describe the distribution of language tags.

$$LE = -\sum_{i=1}^{k} p_i \log_2(p_i), \tag{4}$$

where, k – number of languages, p_i – number of words in language j divided by the total number of words. This metric is 0 for a monolingual corpus and is bounded by equally distributed k languages. Both LE and M-index can be derived from one another.

As a result of applying the above-considered metrics to our data, the values indicated in Table 2 were obtained.

Our dataset shows a high M-Index (58.14%), indicating a balanced distribution of words between the two languages (Table 2). This balance is consistent with the language distribution data presented in Table 1. The high values for both CMI (33.84) and the M-Index confirm a high frequency of code-switching points throughout the corpus.

Furthermore, the token entropy values ($\sim$ 11.5 for both languages) suggest high lexical diversity and an absence of dominant tokens, reflecting a complex and varied vocabulary. When compared to established Hindi-English and Spanish-English datasets (Table 3), our corpus demonstrates a significantly higher CMI. This suggests that our data is not only comparable to but potentially more complex than many existing benchmarks in terms of switching density.

4. N-gram Analysis of Code-Mixing Data

Beyond basic quantitative metrics, we examine the structural characteristics of the code-switching data through character n-grams. The degree of similarity in n-gram distributions between languages directly correlates with the complexity of language identification and sequence labeling tasks. For this analysis, we extracted character n-grams of lengths $n \in \{2, \ldots, 6\}$ from the respective language vocabularies. Figure 2 illustrates the overlap between Ukrainian and Russian n-grams. As is typical for closely related languages, the overlap decreases significantly as the n-gram length increases. A higher overlap probability (e.g., approaching 60%) signifies greater lexical ambiguity, increasing the challenge for computational models to distinguish between the two languages based on sub-word features.

The following analysis highlights the most significant frequency discrepancies between Ukrainian and Russian for 2-grams and 3-grams within our dataset (see Figures 3 and 4). Table 4 presents a comprehensive overview of n-grams ($n \in \{2, \ldots, 6\}$), detailing the total count for each language and the extent of their overlap.

Certain n-gram combinations are orthographically distinct to Ukrainian, as they contain the Cyrillic letters і and ї, which are absent from the Russian alphabet. Many high-frequency n-grams represent characteristic morphemes, such as the Ukrainian -ння *-nnja*, -ськ *-s'k*, and -ти *-ty*, or the Russian пре- *pre-*, -ени(е) *-eni(je)*, and -ть *-t'*. Furthermore, combinations such as -то *-to* (a component of Russian pronouns like то *to*, что *čto*, это *ėto* 'that', 'this') and -ого *-ogo* (the genitive singular masculine ending) highlight differences in frequent word forms. Notably, while -ого *-ogo (RU)/-oho (UK)* exists in both languages, in Ukrainian it appears more frequently in pronominal forms such as його *joho*, нього *n'oho* 'his', 'him', чого *čoho* 'that', нічого *ničoho* 'nothing', and свого *svoho* 'one's own'.

N-gram analysis serves not only to quantify the similarities and differences between Ukrainian and Russian but also as a robust feature for developing token-level language identification models.

5. Code-Switching in Collocations

Beyond static analysis, it is crucial to examine which n-grams emerge at the points of code-switching. Consequently, our study investigates n-gram pat-

Dataset	M-index (%)	I-index	CMI	LE (UK)	LE (RU)
Our Dataset	58.14	7.99	33.84	11.55	11.48

Table 2: Quantitative metrics and language entropy (H) for the Uk-Ru code-switching dataset.

Language pair	CMI index	Source article
English-Bengali	22.48	(Das and Gambäck, 2014)
Dutch-Turkish	22.65	(Nguyen and Doğruöz, 2013)
Spanish-English	22.11	(Mave et al., 2018)
Hindi-English	22.22	(Mave et al., 2018)
Nepali-English	20.32	(Solorio et al., 2014)
Magahi-Hindi-English	51.54	(Rani et al., 2022)

Table 3: Code-mixing index for the different language pair datasets.

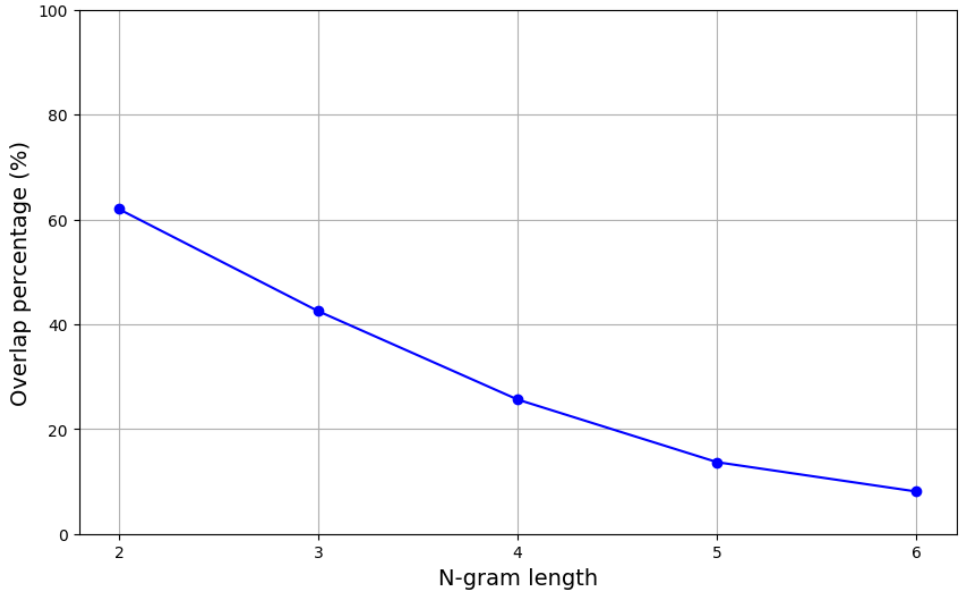

Figure 2: Plot of character N-grams overlap between the Ukrainian and Russian languages, $n \in \{2, \ldots, 6\}$.

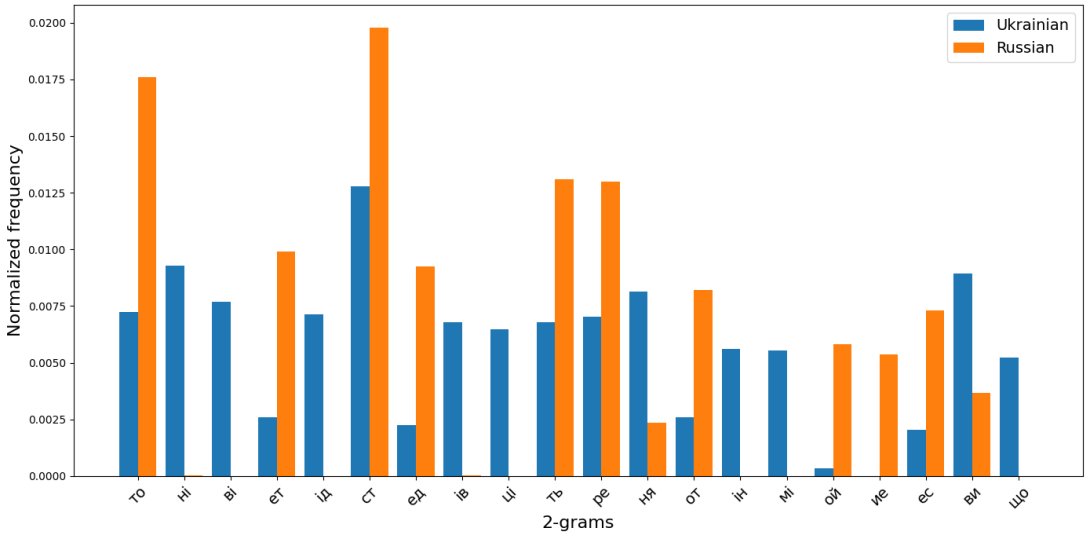

Figure 3: Top-20 2-grams with the greatest frequency discrepancy between Ukrainian and Russian.

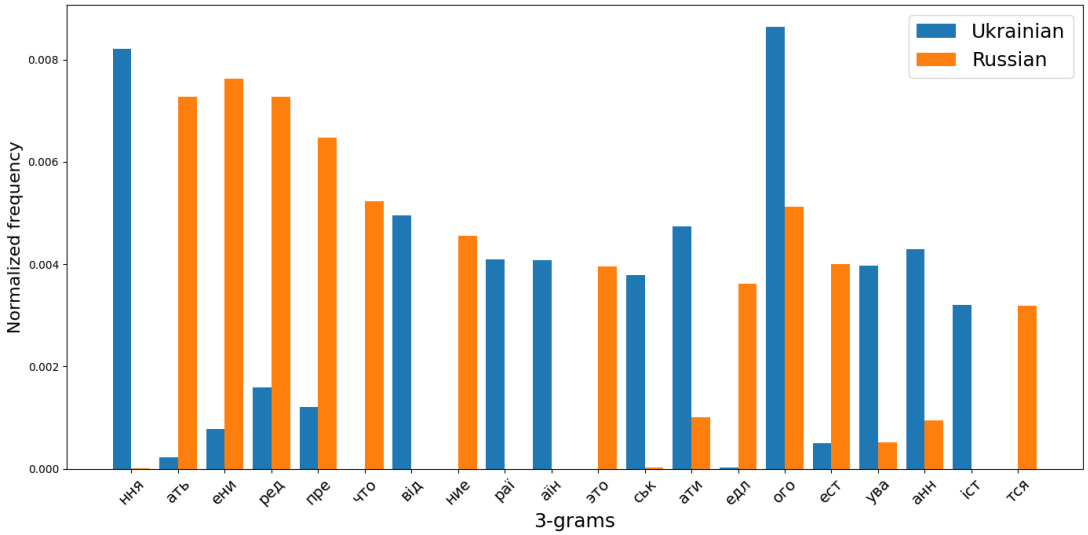

Figure 4: Top-20 3-grams with the greatest frequency discrepancy between Ukrainian and Russian.

N-grams	n=2	n=3	n=4	n=5	n=6
Unique n-grams in Ukrainian	256	2,929	12,248	23,531	28,565
Unique n-grams in Russian	194	2,121	9,438	18,341	22,415
Common n-grams	33	3,730	7,471	6,636	4,478

Table 4: Information about n-grams (n=2..6) of Uk-Ru code-switching dataset.

terns in these transitions, alongside the parts of speech (POS) most frequently involved in switching events (see Figure 5). For morphological analysis, we utilized the *spaCy* model[5].

Most often, the code-switching boundary occurs in syntactically related phrases between an adjective and a noun. This finding aligns with syntactic constraints on code-switching established in earlier linguistics research: the distribution of switching points around nouns and adjectives is consistent with the Equivalence Constraint (Poplack, 1980), which predicts switches at positions where surface syntax is shared across both languages. We analyzed 150 such collocations of the ADJ+NOUN type manually. Most of them are cases where a speaker inserts one or more words in another language. Almost all collocations under consideration are syntactically related, switching can occur in both directions.

5.1. Lexical Code-Switching

This section presents examples of lexical code-switching, where individual words from one language are inserted into an utterance otherwise belonging to the other. The examples are grouped by the direction of the switch: from Ukrainian to Russian ($uk \Rightarrow ru$) and vice versa ($ru \Rightarrow uk$).

[5] https://spacy.io/models/uk

5.1.1. Ukrainian to Russian switching (uk=>ru)

(1) більшу часть
bigger.ACC.F(UK) part.ACC.F(RU)
'the bigger part'

(2) політичний кризис
political.NOM.M(UK) crisis.NOM.M(RU)
'political crisis'

(3) ганебна возня
shameful.NOM.F(UK) fuss.NOM.F(RU)
'shameful fuss'

5.1.2. Russian to Ukrainian switching (ru=>uk)

(4) следующий рік
next.NOM.M(RU) year.NOM.M(UK)
'next year'

(5) остальних продуктів
other.GEN.PL(RU) product.GEN.PL(UK)
'of the other products'

(6) второго зауваження
second.GEN.N(RU) remark.GEN.N(UK)
'of the second remark'

5.2. Idiomatic expressions and calques

Some expressions combining Ukrainian and Russian elements may represent calques of idioms

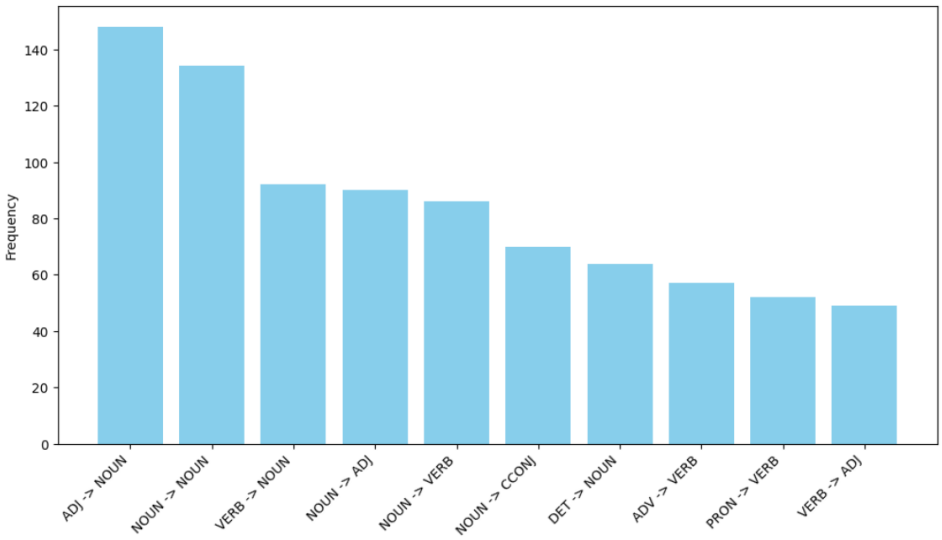

Figure 5: POS switching distribution for the code-switch boundaries.

or frequent collocations. The most common are Ukrainianized calques of Russian expressions:

(7) поетичного слогу
poetic.GEN.M style.GEN.M
'of poetic style'
(UK) (UK)

[calque of RU: поэтического слога; expected UK: поетичного стилю]

(8) железной дорозі
iron.DAT.F road.DAT.F
'railway'
(RU) (UK)

[calque of RU: железной дороге; expected UK: залізниці]

(9) борзими щенками
borzoi.INSTR.PL puppy.INSTR.PL
'with borzoi puppies'
(UK) (RU)

[calque of RU idiom from Gogol's "The Government Inspector": брать взятки борзыми щенками,
lit. 'take bribes with greyhound puppies', 'take bribes in kind rather than money']

5.3. Morphological Code-Switching

Beyond purely lexical switching, morphological patterns reveal more subtle forms of code-mixing where shared lexemes take grammatical forms from different languages. Such patterns are predicted by the Matrix Language Frame model (Myers-Scotton, 1993), which posits that inflectional morphology tends to follow the dominant (matrix) language of the speaker, even when lexical items are borrowed from the other language.

5.3.1. Ukrainian or shared lexemes with Russian morphology

The following collocation could be considered fully Ukrainian, except for the Russian grammatical form:

(10) місцевих бюджетов
local.GEN.PL(UK) budget.GEN.PL(RU,-ов)
'of local budgets'
[expected UK: бюджетів -ів]

In some cases, non-standard Ukrainian endings coincide with standard Russian ones. Such cases may be considered colloquial Ukrainian rather than actual language mixing. The most common cases of this type are:

- **Uncontracted adjective forms**, which may be either Russian or dialectal Ukrainian (northern dialects, see (Zales'kyj and Matviyas, 2001), p. III, map No. 36):

(11) уникальную турботу
unique.ACC.F(RU,-ую) care.ACC.F(UK)
'unique care'
[expected UK: унікальну -у]

(12) новую частину
new.ACC.F(RU,-ую) part.ACC.F(UK)
'new part'
[expected UK: нову -у]

- **Genitive masculine endings** -а , which may be either Russian or colloquial Ukrainian. In Russian, -а is the only possible ending in most

cases (except for a limited number of words that can also have the -y *-u* ending in partitive meaning, but such cases are not considered in our data), while in Ukrainian, part of the nouns have the ending -a and part -y, depending on semantics. This literary norm is often not strictly observed in spoken Ukrainian.

(13) другого етапа
second.GEN.M(UK) stage.GEN.M(RU,-a)
'of the second stage'
[expected UK: етапу -y]

(14) основного капитала
main.GEN.M(UK) capital.GEN.M(RU,-a)
'of the main capital'
[expected UK: капіталу -y]

5.3.2. Russian or shared lexemes with Ukrainian morphology

The following collocations could be considered fully Russian, except for the Ukrainian grammatical form:

(15) досадна ошибка
regrettable.NOM.F(MIX) mistake.NOM.F(RU)
'regrettable mistake'
[expected RU: досадная -ая]

(16) игорного бізнесу
gambling.GEN.M(RU) business.GEN.M(UK)
'of gambling business'
[expected RU: бизнеса -a]

5.3.3. Bidirectional morphological switching

Some collocations can be morphologically normalized, both towards Ukrainian and Russian:

(17) новой редакції
new.GEN.F(RU) edition.GEN.F(UK)
'of the new edition'
[normalized UK: нової редакції]
[normalized RU: новой редакции]

5.4. Orthographic Code-Switching

A controversial issue in annotation is whether cases where the difference between Ukrainian and Russian is only orthographical should be marked as code-switching, since it does not come from the original speaker but is the decision of the transcriber. It can be assumed that the speaker's phonetics influenced the transcriber in such cases, so perhaps they should be considered as well.

5.4.1. Ukrainian phrases with shared words in Russian orthography

Collocations that could be Ukrainian but contain elements of Russian orthography:

(18) слідчої комиссии
investigative.GEN.F(UK)

commission.GEN.F(RU|RU.ORTH?)
'of the investigative commission'
[UK spelling: комісії]

(19) транспортных засобів
transport.GEN.PL(RU|RU.ORTH?)

vehicle.GEN.PL(UK)
'of transport vehicles'
[UK spelling: транспортних]

5.4.2. Russian phrases with shared words in Ukrainian orthography

Collocations that could be Russian but contain elements of Ukrainian orthography:

(20) Огромный дефіцит
huge.NOM.M(RU)

deficit.NOM.M(UK|UK.ORTH?)
'huge deficit' [RU spelling: дефицит]

(21) номінальний прирост
nominal.NOM.M(UK|UK.ORTH?)

increase.NOM.M(RU)
'nominal increase' [RU spelling: номинальный]

5.5. Syntactic-Level Code-Switching

Clearly, in some cases, code-switching cannot be described at the token level only.

5.5.1. Ambiguous prepositions

The combination ADP+NOUN is also frequent, but not always informative, because most common prepositions in Ukrainian and Russian are orthographically identical: на *na* 'on', в *v* 'in', у *u* 'in/near', до *do* 'to', за *za* 'by', про *pro* 'about', для *dl'a* 'for', etc. Determining the language of such prepositions is primarily a matter of annotation principles: whether to consider the preposition at the code-switching boundary as part of the grammatical form of the noun and attribute it the language of the noun, or to consider it a separate item, and in this case the boundary can be between the preposition and the noun.

Examples of such questionable case:

(22) їм плескали в ладони
they.DAT clap.PST in palm.ACC.PL
'[people] clapped for them'
(UK) (UK) (UK/RU?) (RU)

Verbal government as a disambiguation cue

In some cases, syntactic patterns can help determine the language of ambiguous prepositions:

(23) Присягаю на верность Украине
swear.1SG on loyalty Ukraine.DAT
'I swear loyalty to Ukraine'
(UK) (UK/RU?) (RU) (RU)

In this example, although the preposition на 'on' is orthographically identical in Ukrainian and Russian, it belongs to the Ukrainian verbal government pattern присягати на + Accusative (swear on + Accusative), whereas Russian uses a different construction клясться в + Locative (swear in + Locative). This suggests that despite the ambiguous preposition, the syntactic structure follows Ukrainian grammar, with only the lexical items верность 'loyalty' and Украине 'Ukraine' being Russian.

Clear cross-language preposition-noun combinations

There are cases when a preposition and a noun definitely belong to different languages, e.g.:

(24) у кавычках
in quotation.mark.LOC.PL
'in quotation marks'
(UK) (RU)

(25) к діям
to action.DAT.PL
'to actions'
(RU) (UK)

5.5.2. Context-dependent ambiguity

Finally, there are collocations where one of the words is orthographically completely identical in the two languages, and the code-switching can only be determined based on the wider context, not always with absolute certainty:

(26) з секретного соглашения
from secret.GEN.N agreement.GEN.N
'from the secret agreement'
(UK) (UK/RU?) (RU)

Full context: Я (UK) хотів (UK) би (UK) звернути (UK) увагу (UK), що (UK) все (UK) це (UK) почалося (UK) з (UK) секретного (UK/RU?) соглашения (RU) в (RU) отношении (RU) уровня (RU) заработной (RU) платы (RU) Тимошенко (RU)
'I would like to draw attention that all this began with a secret agreement regarding Tymoshenko's salary level'

The collocation analysis shows that taking syntactic relations and idiomatic expressions into account is a promising approach for further study of code-switching. The results of the POS switching distribution (Figure 5) provide insights into the syntactic structures where code-switching is most likely to occur. These results highlight that code-switching is not random but rather occurs in specific syntactic environments, particularly around nouns and their modifiers or connected verbs. This is consistent with established findings that certain parts of speech, particularly nouns and their modifiers, are especially susceptible to code-switching. Understanding these patterns can provide deeper insights into the linguistic mechanisms of code-switching in the Ukrainian-Russian bilingual context.

6. Conclusion

In this paper, we evaluated a comprehensive set of token-level metrics on a Ukrainian-Russian code-switching dataset. Our results demonstrate that this corpus exhibits a high switching density, surpassing several established benchmarks in the field. Through a detailed n-gram analysis ($n \in \{1, \ldots, 6\}$), we quantified the degree of cross-linguistic overlap and identified language-specific sub-word features that are critical for robust language identification.

Furthermore, our investigation into the morphological patterns at switching points reveals that code-switching is not random but follows specific part-of-speech distributions, providing empirical, corpus-based support for established theoretical frameworks in the linguistic literature that posit morphosyntactic structure as a key constraint on code-switching behavior (Poplack, 1980; Myers-Scotton, 1993; Muysken, 2000). These findings suggest that integrating morphological and syntactic features can significantly enhance the performance of automated code-switching detection systems. This work provides both a validated dataset and a methodological foundation for future research in Slavic-centric multilingual NLP.

Acknowledgments

We would like to thank the reviewers for their time and effort in reviewing this manuscript. We sincerely appreciate their valuable comments and suggestions, which greatly helped us improve the quality of the work. This research was partially funded by the Alexander von Humboldt Foundation, and this work received support from the COST Action CA21167 'UniDive'[6] (European Cooperation in Science and Technology). The authors are also grateful to Friedrich Schiller University Jena for providing the research facilities and support that made this work possible.

[6] https://unidive.lisn.upsaclay.fr/

7. Bibliographical References

References

B. Barnett, E. Codo, E. Eppler, M. Forcadell, P. Gardner-Chloros, R. van Hout, M. Moyer, M. Torras, M. Turell, M. Sebba, M. Starren, and S. Wensink. 2000. The LIDES coding manual - A document for preparing and analyzing language interaction data. Version 1.1, July 1999. *International Journal of Bilingualism*, 4(2):131–270.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, Sascha Diwersy, Maria Gavriilidou, Ruben van Heusden, Mikel Iruskieta, Neeme Kahusk, Anna Kryvenko, Noémi Ligeti-Nagy, Carmen Magariños, Martin Mölder, Costanza Navarretta, Kiril Simov, Lars Magne Tungland, Jouni Tuominen, John Vidler, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, and Darja Fišer. 2024. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

G. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio. 2017. Metrics for modeling code-switching across corpora. In *Proceedings of Interspeech 2017*, pages 67–71, Stockholm, Sweden. ISCA.

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 12–20, Austin, Texas. Association for Computational Linguistics.

Nikolay Hakimov. 2021. *Explaining Russian-German code-mixing*. Number 3 in Contact and Multilingualism. Language Science Press, Berlin.

Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova, and Ruprecht von Waldenfels. 2023. The parliamentary code-switching corpus: Bilingualism in the Ukrainian parliament in the 1990s-2020s. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 79–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Olha Kanishcheva, Maria Shvedova, Liudmyla Dyka, and Kristina Husenko. 2026. Study of language identification task on the token level for Ukrainian-Russian code-switching dataset. *Northern European Journal of Language Technology*, 12(1).

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge University Press, Cambridge.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press, Oxford.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7–8):581–618.

Priya Rani, John P. McCrae, and Theodorus Fransen. 2022. MHE: Code-mixed corpora for similar language identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433, Marseille,

France. European Language Resources Association.

Maria Shvedova and Arsenii Lukashevskyi. 2025. Case choice in Ukrainian vocative expressions: A study of parliamentary transcripts (1990–2024) annotated with Universal Dependencies. In *Grammar and Corpora: 10th International Conference, Book of Abstracts*, pages 106–108, Riga. University of Latvia Press.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Antin Zales'kyj and Ivan Matviyas, editors. 2001. *Atlas of the Ukrainian Language*, volume 3. Naukova Dumka, Kyiv. 206 maps.