Study of Language Identification Task on the Token Level for Ukrainian-Russian Code-Switching Dataset

Olha Kanishcheva, Heidelberg University, Germany, SET University, Ukraine
o.kanishcheva@setuniversity.edu.ua

Maria Shvedova, National Technical University "Kharkiv Polytechnic Institute", Ukraine
mariia.shvedova@khpi.edu.ua

Liudmyla Dyka, National University of Kyiv-Mohyla Academy, Ukraine dykalv@ukma.edu.ua

Kristina Husenko, University of Helsinki, Finland kristina.husenko@helsinki.fi

Abstract This paper presents experiments on language identification for a Ukrainian-Russian code-switching dataset. Code-switching, a common phenomenon in multilingual societies, presents significant challenges for natural language processing. This study discusses various issues encountered during dataset creation, emphasizing the complexity of accurately annotating code-switching text. The study describes cases where identifying the language of individual tokens in sentences that switch between Ukrainian and Russian proves difficult even for human annotators. The relatedness of the languages and the use of Cyrillic in both orthographic systems complicate the task, leading to many cases where words are spelled identically despite clear phonetic differences between the languages that are not reflected in writing. The study explores different models and libraries for language identification on the token level. Experimental results suggest that BERT shows promising performance; however, other models, such as CRFs with n-grams, Char-level BiLSTM, and Word-level Neural Networks, are also promising for this task. This research contributes to the development of language processing technologies for multilingual contexts, with potential applications in sentiment analysis, information retrieval, and social media monitoring.

1 Introduction

In multilingual societies, code-switching or code-mixing, the alternating use of multiple languages within discourse, presents a unique challenge for natural language processing (NLP). Understanding and identifying languages within code-switched texts is crucial for various NLP applications such as sentiment analysis, machine translation, information retrieval, among others. Ukrainian is a notable example of a language frequently involved in code-switching.

The terms *code-switching* and *code-mixing* are commonly used in the study of the use of two or more languages in a single discourse. Researchers adopt different approaches to distinguishing between these phenomena. One approach focuses on structural differences: *code-mixing* typically involves seamless mixing of linguistic elements within a single utterance, while *code-switching* involves distinct switches between languages at identifiable points within a conversation. In some classifications, the term *code-switching* encompasses different formal types of switching, which are subdivided into intra-sentential, intra-word, and inter-sentential types (Schmidt, 2014). Another approach distinguishes between these terms based on the meaning that speakers attribute to such events, particularly whether the switches are intentional (*code-switching*) or unintentional (*code-mixing*) (Hakimov, 2021). This study focuses on intra-sentential code-switching, as our corpus consists of isolated sentences. Therefore, in this paper, we will use the term *code-switching* as an umbrella term, referring to both code-switching and code-mixing without making a distinction between them.

Due to variations in spelling and grammar, code-switching in social media material presents significant challenges for natural language processing (Mave et al., 2018a). However, existing language identification models often struggle to accurately identify Ukrainian segments within code-switching corpora.

In this context, we conduct a comprehensive evaluation of various language identification libraries on a code-switching dataset. Language identification is a fundamental NLP task aimed at determining the lan-

https://doi.org/10.3384/nejlt.2000-1533.2026.5725

guage of a given text segment. It plays a crucial role in numerous applications, including information retrieval, machine translation, sentiment analysis, and content filtering. Major challenges in this task include the processing of short texts with sparse linguistic information and distinguishing between closely related languages, such as Spanish and Portuguese, or in our case, Ukrainian and Russian.

We hypothesize that contextual language models, such as BERT, can effectively identify language at the token level in Ukrainian-Russian code-switching scenarios due to their ability to capture fine-grained contextual and morphological cues. The goal of this paper is to develop a token language identification model for the Ukrainian-Russian code-switching dataset based on parliamentary transcripts from the Verkhovna Rada (the unicameral parliament of Ukraine). We discuss the development of a specialized dataset, emphasizing the challenges faced during data annotation. The models evaluated for the token classification task include the Conditional Random Fields (CRFs), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), and BERT.

The structure of our article is as follows: in the *Introduction*, we outline the challenge of language identification in Ukrainian-Russian code-switching contexts. Section 2 discusses the previous research in the field and the existing approaches to language identification in code-switching scenarios. In Section 3, we explain how the dataset was compiled and annotated, highlighting key obstacles faced during this process. Section 4 assesses the performance of several language identification tools when applied to our dataset. Section 5 describes the models used for language identification at the lexical level, presents our experimental setup and results, and concludes with an error analysis. Finally, *Conclusion* summarizes the findings and implications of our study.

2 Related Work

Language identification in code-switching datasets has attracted considerable attention in recent years, with researchers exploring various methodologies and approaches to address the unique challenges presented by multilingual data (Winata et al., 2023). In the paper (King et al., 2014), the authors conducted experiments on language identification for similar languages such as Bosnian, Croatian, and Serbian, Indonesian and Malay, Czech and Slovak, Brazilian and European Portuguese, Argentinian and Peninsular Spanish, and American and British English. Furthermore, several benchmarks have been developed to support systematic evaluation of code-switched NLP models, most notably LinCE (Aguilar et al., 2020) and GLUECoS (Khanuja et al., 2020), which provide standardized datasets and tasks for multiple language pairs, fostering comparability and progress in multilingual research.

Previous studies have employed both traditional machine learning techniques and state-of-the-art deep learning models to tackle language identification tasks in code-switching contexts (Jose et al., 2020).

Several studies have focused on utilizing statistical methods such as Conditional Random Fields (CRFs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs) for language identification in code-switching datasets (Hindi-English and Spanish-English datasets) (Mave et al., 2018b; Hidayatullah et al., 2023). These methods often rely on features derived from linguistic characteristics, such as n-grams, part-of-speech tags, and syntactic structures, to distinguish between different languages within a mixed-language dataset.

In recent years, the advent of deep learning has led to the development of neural network-based approaches for language identification in code-switching data. Techniques such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) were used for the Hindi-English data (Joshi and Joshi, 2020), and Transformer-based architectures like BERT have shown promising results in handling the complexities of code-switched text (Malayalam-English, Marathi-English and Indonesian-Javanese-English corpora) (Thara and Poornachandran, 2021; Chavan et al., 2023; Hidayatullah et al., 2023). These models leverage the hierarchical and contextual nature of language to capture intricate patterns and dependencies between languages present in mixed-language utterances.

Furthermore, researchers have explored the use of ensemble learning techniques and transfer learning paradigms to enhance the robustness and generalization capabilities of language identification models for code-switching datasets (Zhang et al., 2018; Aguilar and Solorio, 2020). By combining multiple classifiers or leveraging pre-trained language models, these approaches aim to improve performance across different language pairs and domains.

Overall, the field of language identification in code-switching datasets is characterized by a diverse range of methodologies and a growing body of research aimed at addressing the unique linguistic challenges inherent in multilingual communication.

However, such resources and studies are limited for the Ukrainian language. A few individual efforts can be noted, such as (Pylypenko and Lyudovyk, 2019), which deals with audio analysis, and (Sira et al., 2019), where the goal was limited to rule-based identification of mixed speech.

The Ukrainian linguistic tradition has studied code-switching mainly in the prescriptive context as a deviation from the Ukrainian language norm through interference with the Russian language. Although some recent works have started to analyze different sociolinguistic scenarios of Ukrainian-Russian code-switching and code-mixing more deeply (Mikolchak, 2025). Code-mixed Ukrainian-Russian speech is referred to as *surzhyk* (in the first meaning, *a mixture of rye and any other bread grain that is milled during poor harvests* (Biletskyi-Nosenko, 1966)). Larysa Masenko (Masenko, 1999, 2011), Oleksandr Taranenko (Taranenko, 2007), and other researchers have studied surzhyk in contrast to the standard language, as a violation of the literary norm. Michael Moser raises the question of the possible systematic nature of the surzhyk and demonstrates the high variation within it, which seems to testify against the idea of its regularity (Mozer, 2016). Currently, the project *Hybridisierung von zwei Seiten*[1] is conducting the first systematic corpus-based studies of surzhyk based on a large corpus of recordings of hybridized Ukrainian-Russian spoken language from different regions, which was created at the University of Oldenburg (Hentschel and Taranenko, 2022; Hentschel, 2024). The corpus itself is to be published in the future.

This points to a relative lack of academic research and NLP resources, indicating a need for more focused work on the technical processing of code-switching involving the Ukrainian language. Addressing this gap would provide valuable insights into the linguistic dynamics and sociolinguistic factors influencing code-switching behaviors in Ukrainian-speaking communities. Expanding research in this area could also contribute to developing more comprehensive multilingual resources and tools. These would better reflect the linguistic diversity and complexities of Ukrainian in code-switching studies.

3 Dataset Creation and Annotation

In this study, we work with Ukrainian parliamentary transcripts (1990-2024)[2], focusing on utterances 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). Moreover, Ukrainian parliamentary transcripts provide additional linguistic value, as they are recorded verbatim, preserving features of spoken language, such as hesitations, grammatical errors, and code-switching. Unlike the previously mentioned Oldenburg Corpus, which contains hybridized spoken texts, the corpus of Ukrainian parliamentary transcripts represents standard-oriented speech. The main language is standard Ukrainian, with a minor presence of Russian, which declined annually and virtually disappeared after 2017 (Kanishcheva et al., 2023). However, some sentences from the transcripts still exhibit code-switching, which provides material for creating a bilingual dataset.

We selected sentences that contain a mix of both languages, leveraging this dataset for its rich and authentic instances of bidirectional code-switching between Ukrainian and Russian. A key challenge is that Ukrainian and Russian are closely related Slavic languages that use a similar Cyrillic alphabet, making automatic language identification difficult.

3.1 Data Creation

The dataset consists of individual sentences extracted from the Ukrainian parliamentary transcripts. To focus on mixed-language content, we excluded sentences that were entirely or predominantly in Russian using CleanText.groovy[3]. The remaining sentences were lemmatized with the dictionary-based TagText parser[4], and filtered to retain only those containing more than two out-of-vocabulary words, which typically indicate a mixture of Ukrainian and Russian. A small subset of the selected sentences also included spelling errors or non-dictionary terms. This approach resulted in a dataset of approximately 150,000 tokens. All sentences were tokenized, and each token was manually labeled with its corresponding language. To create a balanced dataset, so that the number of tokens in Ukrainian was close to the number of tokens in Russian, we added the Russian sentences that were previously excluded. Detailed statistics for the final dataset are provided in Table 1.

The annotation process involved assigning each token to one of six categories: *Ukrainian*, *Russian*, *Ukrainian-Russian hybrid words (surzhyk)*, *Numbers*, *Others*, and *Punctuation*. The dataset has been published on Zenodo[5] under the Creative Commons Attribution 4.0 International license, and the label distribution is summarized in Table 2.

The dataset was annotated by three native bilingual speakers of Ukrainian and Russian. Initially, we underestimated the complexity of the task and adopted a single-annotator approach: 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

[1] https://uol.de/slavistik/forschung/sprachwissenschaft/hybridisierung-von-zwei-seiten

[2] https://www.rada.gov.ua/meeting/stenogr

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

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

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

Metric	**Value**
Sentences	8,849
Average sentence length (characters)	144.84
Average sentence length (tokens)	23.49
Minimum sentence length (characters)	1
Minimum sentence length (tokens)	1
Maximum sentence length (characters)	1,401
Maximum sentence length (tokens)	232
Words	218,809
Unique words	19,304

Table 1: Statistics for the Ukrainian-Russian code-switching dataset.

task was more nuanced than anticipated, particularly in the areas discussed in Section 3.3: distinguishing between code-switching and lexical borrowings, identifying boundaries between dialectal forms and Russian influence, handling syntactic calques, and resolving ambiguous cases of words with similar spelling in both languages. 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. This iterative process led to the systematized annotation principles described in Section 3.3.

Analysis of the tag distribution in the dataset (Table 2) reveals a balanced representation of Ukrainian and Russian, with 41.85% and 37.65%, respectively. However, Ukrainian-Russian hybrid tokens account for only 0.29%, leading us to exclude the *MIX* category from our classification model.

Since one of our goals is to develop a dataset for closely related language identification, understanding the similarities between these languages is crucial. Both Ukrainian and Russian belong to the Slavic language family and share geographic regions, resulting in extensive contact and many shared linguistic features. Therefore, we first examined the lexical overlap between the two languages. This analysis helps us better understand the complexity of the identification task and the challenges presented by code-switching.

To calculate lexical overlap, we first lemmatized[6] all words and removed duplicates within each language. We then measured the overlap between Ukrainian and Russian tokens in our dataset (Table 3).

Approximately 7.6% of Ukrainian lemmas overlap with Russian lemmas in our dataset, which increases the difficulty of the language identification task.

The distribution of languages per sentence is shown in Figure 1.

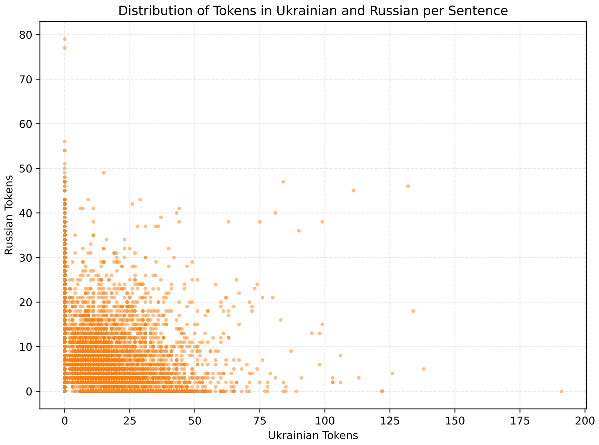

Figure 1: Distribution of tokens in Ukrainian and Russian languages per sentence.

Manual token-level language annotation of Ukrainian-Russian code-switching data by language presents several challenges. Below, we describe our approach to addressing them.

3.2 Transcription Errors

The verbatim recording of speech in the transcripts contains numerous typographical errors due to the speed of transcription. Spelling and grammatical errors in our data can present a problem for automatic language identification in transcripts. When transcribing sentences with switching, the transcriber does not always switch the keyboard layout in time, writes in transliteration, and the spelling does not match the language of the sentence.

However, some transcription errors may hold research value, as they can provide insights into pronunciation. For instance, in some cases, the transcript records the speaker's self-correction of their pronunciation. Nevertheless, a transcript is not a phonetic record and should not be considered a reliable source for determining pronunciation.

In the version of the data intended for training the model, such obvious transcription errors have been corrected, and these corrections have been annotated.

3.3 Challenging Cases of Manual Language Detection

The analyzed Ukrainian-Russian mixed sentences contain fragments that switch from Ukrainian to Russian or from Russian to Ukrainian. Most often, the switching point can be clearly identified and is associated with

[6] https://github.com/brown-uk/nlp

Labels	Description	Count	%
UK	Ukrainian words	91 592	41.85
RU	Russian words	82 375	37.65
MIX	Ukrainian–Russian hybrid words (surzhyk)	652	0.29
NUM	Numbers	2123	0.97
OTH	Dialects, other languages, etc.	234	0.10
PUNCT	Punctuation	41 832	19.11

Table 2: Dataset statistics for the language pair Ukr-Rus.

Token Statistics	Count
Sentences	8,849
Lexical overlap between Ukrainian and Russian tokens	742
Total lemmatized Ukrainian tokens	10,434
Ukrainian tokens after removing duplicates	8,441
Total lemmatized Russian tokens	9,692
Russian tokens after removing duplicates	7,699

Table 3: Token-level statistics and lexical overlap in the Ukrainian-Russian dataset.

quotations from laws and documents or the use of official names or untranslated phraseology:

(1) *...**поданный Кабинетом Министров на рассмотрение во втором чтении** проект "Закону України про здійснення платежів на території України...*

*...**submitted by the Cabinet of Ministers of Ukraine for consideration in the second reading** of the draft Law of Ukraine on making payments on the territory of Ukraine...*
(Hereinafter, the Ukrainian text is in italic, the Russian text is in bold).

However, there are also numerous cases where language identification of individual tokens can be problematic from a linguistic perspective. Each type of case requires separate attention. Below, we describe the annotation principles developed for these challenging cases.

3.3.1 Words with Similar Spelling in Both Languages

Since the languages are closely related, there are many words that look the same in writing. Such words are labeled as Ukrainian or Russian based on context. Difficulties arise when such ambiguous words are used in mixed sentences, especially at code-switching boundaries. It is not obvious to which language they should be assigned.

(2) *Я (uk|ru?) хочу (uk|ru?) внести (uk|ru?) пропозицию, **Владимир Михайлович, это** пропозиция про врахування **представления следующих изменений***

I (uk|ru?) want to (uk|ru?) make (uk|ru?) a proposal, **Vladimir Mikhailovich, it is** *a proposal for the consideration* **of the following changes to the submission**

(3) *В ході діяльності ... комісії вияснилось (ru|uk?), **что одна и та же особа представляла различные компании.***

*In the course of the commission's work, it turned out (ru|uk?) **that the same person represented different companies.***

In our dataset, such tokens were tagged as either Ukrainian or Russian. Introducing a separate 'ambiguous' tag for them could be considered. However, the introduction of such a category would raise additional questions. Ambiguity is sometimes present even when a word contains unambiguously Russian or Ukrainian letters, as the transcriber works quickly and is not always able to determine in time which spelling should be used and switch to the appropriate layout. In cases of doubt, we tagged the tokens as Ukrainian, since Ukrainian predominates in the entire corpus of Ukrainian parliament transcripts. Furthermore, such ambiguous cases are rare and do not constitute a systematic phenomenon in the dataset.

3.3.2 Hybrid Words and Morphological Adaptations

It can be assumed that, in many cases, the few Russian words in the Ukrainian text were actually pronounced with Ukrainian phonetics and thus should be classified as mixed, but unfortunately, the transcript does not reflect this. This is confirmed by audio recordings (which are not available online for all transcripts) and by cases where words that seem to be Russian in writing are inflected later within the same text according to the Ukrainian morphological model. This leads to inconsistent annotation, as this lexeme receives the *ru* tag in the first instance and the *mix* tag in the second instance.

(4) ***Огромный*** *дефіцит бюджету, огромна (mix) дира (mix)*

__Huge__ budget deficit, huge (mix) hole (mix)

In Example (4), which is predominantly Ukrainian, the Russian word огромный *ogromnyj* 'huge' occurs twice, in the second instance with the Ukrainian feminine singular ending *-a*.

Hybridized words that have, for example, a Russian root and a Ukrainian ending, as in Example (4), are marked with the *mix* tag. Similarly, Ukrainian words within a Russian phrase, with adaptation to the Russian morphological system, are marked with the *mix* tag (5, 6).

(5) *звернулись . . .* ***с открытой*** *заявой (mix)*

have come forward ... ***with an open*** *statement (mix)*

Language mixing within a word can occur on the boundary of language switching:

(6) *кінець квартала, конець (mix) півріччя, конець (mix)* ***года и так далее***

end of the quarter, end (mix) of the half-year, end (mix) ***of the year, and so on***

3.3.3 Syntactic Borrowings and Calques

Ukrainian-Russian hybridization (surzhyk) can occur not only at the token level but also at the phrase level. Annotating a language only at the token level is not really appropriate for such cases. These include constructions based on Ukrainian words but following the Russian syntactic model, calques of Russian phraseology, paraphrases of precedent texts, the use of Russian discourse elements in Ukrainian texts, etc.

(7) *я дуже дякую* ***вас***

I thank you very much

In Example (7), the word вас *vas* 'you' is the correct Ukrainian (and Russian) accusative case of the second person plural personal pronoun. However, the proper Ukrainian form here should be the dative case вам *vam*; the accusative case is borrowed from a similar Russian phrase: я благодарю **вас** *ja blagodarju* ***vas*** (Acc.) 'I thank you'.

3.3.4 Boundaries Between Borrowings and Code-Switching

Another complication is that it is sometimes difficult to define the boundary between code-switching and borrowing, since when languages come into close contact, some items from another language gradually move into the register of borrowings. In Example (8), we can see the recent borrowing from Russian: the phrase русский мир *russkij mir* 'Russian world' is regularly used in Ukrainian without translation. There is wordplay: the Russian word мир *mir* 'world' sounds similar to 'peace' in Ukrainian and Russian.

(8) *ідеї так званого* ***"русского мира"****, який насправді є* ***миром*** *не* ***русским****, є* ***миром*** *кремлівським, є* ***миром*** *людей, які не хочуть миру*

the ideas of the so-called ***"Russian world"*** *which is not really a* ***Russian world****, it is a Kremlin* ***world****, it is a* ***world*** *of people who do not want peace*

Similarly, the speech of Russian-speaking Ukrainians contains numerous borrowings from Ukrainian (Shvedova, 2016), such as the emphatic particle та *ta* (9).

(9) ***Та берите вы пробы себе***

Just take the samples for yourselves

3.3.5 Boundaries Between Colloquial Forms and Code-Switching

Ukrainian dialectal, colloquial, and rare forms and lexemes that may resemble Russian but are not the result of Russian influence are tagged as Ukrainian. Sometimes, it is not easy to determine the origin of such words and forms unambiguously. What appears to be surzhyk may actually be a native Ukrainian word, but the influence of Russian may still affect its frequency and collocations.

The main criterion for annotating a word as Ukrainian was its presence in the academic dictionary of the Ukrainian language (Bilodid, 1970–1980). However, this dictionary was published more than 50 years ago in Soviet times and contains some words and collocations that modern Ukrainian philologists and editors might classify as surzhyk (10, 11).

(10) *це має відношення до свободи слова*

It is related to freedom of speech (lit. *It has relation to freedom of speech*).

The example may now be perceived as a Russian calque; cf. Russian: *это имеет отношение к свободе* *слова* (lit. *It has relation to freedom of speech*). In modern Ukrainian, the standard variant is *це стосується свободи слова* (lit. *It relates to freedom of speech*).

(11) *Ви сама корумпована партія*

You are the most corrupt party

The analytical adjectival superlative is a regular grammatical form in Russian; cf. Russian: *Вы самая коррумпированная партия.* In Ukrainian, this variant is less frequent and more colloquial; the standard form is synthetic with the prefix най- *naj-*: *Ви найкорумпованіша партія.*

For words that are not in the dictionary, we also used corpus frequency as a criterion (Shvedova et al., 2017–2025). This explains the inconsistency in the annotation of active participles: although modern prescriptive Ukrainian linguistics consider these forms uncharacteristic of the Ukrainian language and a result of Russian influence, frequently used participles (e.g. існуючий *isnujučyj* 'existing', діючий *dijučyj* 'acting') are tagged as Ukrainian, while rarely used ones (e.g. вишестоящий *vyšestojaščyj* 'superior', вилізший *vylizšyj* 'came out') are tagged as *mixed*. There are also other units that are not in the dictionary, but have hundreds of occurrences in the corpus, e.g. (12).

(12) *різні трактовки освіти*

different understandings of education

The word трактовка *traktovka* 'understanding' appears in the Russian dictionary and is not found in the Ukrainian dictionary, where there is only the variant трактування *traktuvannja*. However, in Ukrainian texts it appears with significant frequency (908 findings in GRAC.v.18 from the early 20th century to the present).

3.3.6 Boundaries Between Dialectal Forms and Code-Switching

Another problem is the distinction between surzhyk and Ukrainian dialectal forms that may resemble Russian but are not the result of Russian influence. In (13), it is likely that the transcriber did not recognize the Ukrainian prefix од- *od-* (a less common variant of the prefix від- *vid-*) and mistakenly transcribed it as a Russian word with the prefix от- *ot-*.

(13) *ви їм* ***отдали*** *пільги*

You ***gave*** *them benefits*

Basically, separate Ukrainian dialectal words and forms that appear in the transcripts (14) are labeled as Ukrainian without any special tags.

(14) *не зробе того невеликого вкладу*

will not make that small contribution

In this example, the dialectal verb form of the third person singular зробе *zrobe* 'will make' is used instead of standard зробить *zrobyt'*.

We labeled such interspersed dialectal forms as Ukrainian, in contrast to the larger fragment from the parliamentary transcript in the Transcarpathian dialect, which was labeled *(oth)* since in this case there is an obvious instance of code-switching (15).

(15) *Наші селяни кажуть,* ***ож они русинами були дотепер, той по своєму хотят быты і хотять ся соєдинити из Украïнов (oth).***

Our peasants say ***that they have been Rusyns until now, but they want to be in their own way and want to unite with Ukraine (oth).*** (In this example, we use bold to indicate the Transcarpathian dialect)

3.3.7 Annotation of Neologisms and Ad Hoc Word Formation

The occasional words created on the basis of the Ukrainian language are tagged as Ukrainian (16). In cases where there is an influence of the Russian language, such new words are tagged as *mixed* (17).

(16) *там є застерега (uk)* (instead of *застереження*)

there is a warning

(17) *конкурувати на загальних засновах (mix)* (instead of standard Ukrainian *на загальних засадах*, the hybridized word was created ocasionally under the influence of the Russian *на общих основаниях*)

to compete on a common basis

Thus, language annotation at the token level in Russian-Ukrainian code-switching sentences is a challenging task that has a number of problems even for human annotation. We developed specific annotation principles described in this section to address these challenges.

4 Methods

4.1 Baseline Systems

Various approaches are employed for language identification. Traditional methods often use statistical models analyzing character n-grams or word frequencies. Machine learning techniques, including Support Vector Machines and Naive Bayes, have been successful, with the emergence of deep learning models like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) showing promise (Burchell et al., 2023).

Several tools and libraries facilitate language identification tasks, such as pycld2,[7] Fasttext,[8] langid.py,[9] Spacy,[10] CLD3,[11] Langdetect,[12] Lingua[13] and GlotLID[14]. These tools provide efficient ways to implement language identification algorithms, making them accessible for developers and researchers alike.

Detailed studies of different libraries for language identification are presented in this resource[15]. However, the libraries were compared at the sentence level on the tatoeba-sentences-2021-06-05 data[16] rather than at the token level.

Nevertheless, all these modules work rather poorly with short sentences and consequently with language detection at the token level (Goswami et al., 2020; Mario, 2021).

To solve the task of language identification at the token level, we first tested existing language identification libraries to find the most suitable one for processing Ukrainian-Russian code-switching texts. These libraries received tokenized sentences as input and predicted the language for each token. Only tokens that were unambiguously annotated as either Ukrainian or Russian were included in the evaluation; mixed or ambiguous tokens were excluded. We treated the task as a binary classification problem and evaluated library performance using standard metrics: Accuracy, macro-averaged Precision, Recall, and F1-score, to compare how well they handle code-switching. We report multiple metrics to capture complementary aspects of model performance, as some libraries tend to overpredict the majority language, which can lead to skewed Accuracy values. Detailed metric values are provided in Table 4.

The results presented in Table 4 show that libraries such as Spacy, pycld2, and Lingua can be used for language identification of tokens in a code-switching dataset, but the accuracy of these libraries is still quite low. The low results are primarily due to a lack of context, as most libraries work with sentences rather than individual words. However, one of the latest libraries, GlotLID (Kargaran et al., 2023), has shown very good results in identifying the Ukrainian language at the token level.

4.2 Model Architecture and Training Details

In this section, we will briefly describe the models that we used in our experiments of token-level identification and present our experiment results. The data were not preprocessed and were split into training and test datasets with distributions of 80% and 20%, respectively.

Previous studies (Rani et al., 2022; Mave et al., 2018a; Hidayatullah et al., 2023; Thara and Poornachandran, 2021) have demonstrated the effectiveness of various machine learning approaches for token-level language identification in code-switched datasets. However, no single method consistently outperforms others across all language pairs and scenarios.

For instance, Rani et al. (2022) introduced UDLDI, an unsupervised method combining sentence embeddings with clustering, achieving an F_1 score of 0.89 on a Magahi-Hindi-English dataset-outperforming LSTM and CNN baselines. Mave et al. (2018a) evaluated CRF, BiLSTM, and LSTM models on Hindi-English data and found that CRF with part-of-speech (PoS) features yielded the best performance (F_1 = 0.96–0.98). Similarly, Hidayatullah et al. (2023) and Thara and Poornachandran (2021) explored combinations of CRF, BiLSTM, mBERT, BERT, and ELECTRA, reporting varied outcomes; BERT models frequently performed well, although CRF remained competitive.

Based on these findings, we evaluated several modeling approaches on our code-mixed Ukrainian-Russian dataset: Conditional Random Fields (CRF), Bidirectional LSTM (BiLSTM), a Word-level Neural Network, and BERT.

CRF is a probabilistic graphical model effective for sequence labeling tasks, capable of capturing both local and global dependencies (Mave et al., 2018b; Hidayatullah et al., 2023). LSTM networks are designed to model sequences and long-term dependencies (10.1162/neco.1997.9.8.1735). However, in this study, we also employ word-level neural classifiers based on the LSTM architecture to establish a baseline for classification using only lexical embeddings without broader context. BiLSTM further extends this by processing input in both directions, enhancing contextual representation (Graves and Schmidhuber, 2005). Finally, BERT, a Transformer-based model, leverages deep contextual embeddings and has achieved state-of-the-art results across numerous NLP tasks, including sequence labeling (Devlin et al., 2019).

In our experiments, we implemented and tested these architectures using various feature configurations. We selected the embedding dimension, hidden units, and dropout rates based on prior work on token-level language identification and preliminary experiments on our dataset. The CRF model was trained

[7] https://pypi.org/project/pycld2/

[8] https://huggingface.co/facebook/fasttext-language-identification

[9] https://pypi.org/project/py3langid/

[10] https://spacy.io/usage/models

[11] https://docs.ropensci.org/cld3/

[12] https://pypi.org/project/langdetect/

[13] https://github.com/pemistahl/lingua

[14] https://github.com/cisnlp/GlotLID

[15] https://modelpredict.com/language-identification-survey#reported-metrics

[16] https://tatoeba.org/en

Libraries	Supported languages	Accuracy	Precision	Recall	F_1
pycld2	83	0.75	0.38	0.50	0.43
Fasttext	176	0.16	0.08	0.22	0.12
langid	97	0.43	0.06	0.04	0.05
Spacy	23	0.75	0.38	0.50	0.43
CLD3	107	0.49	0.02	0.01	0.01
Langdetect	97	0.63	0.14	0.11	0.12
Lingua	75	0.75	0.38	0.50	0.43
GlotLID	1665	0.76	0.82	0.77	0.75

Table 4: Evaluation of different language identification libraries on code-switching dataset (tokens that were clearly defined as Ukrainian or Russian were taken for evaluation).

on full token sequences with features including token length, case, character type, and character bigrams. It was optimized with L1 and L2 regularization (c1 = 1.0, c2 = 0.001).

The word-level neural networks were trained on isolated tokens to evaluate the effectiveness of lexical features independently of sentence-level context. We experimented with both randomly initialized embeddings and pre-trained Ukrainian embeddings. The architecture consisted of an embedding layer (embedding dimension = 100), followed by one or two stacked bidirectional LSTM layers with 64 hidden units each, a dropout layer (rate = 0.5) for regularization, and a softmax output layer for binary classification. The models were trained for 6 epochs using a batch size of 32 and the Adam optimizer. While this architecture utilizes LSTM layers, for isolated tokens (sequence length $L = 1$), it effectively functions as a feed-forward neural network. This model serves as a lexical baseline to be compared with more complex sequential approaches.

For context-aware modeling, we fine-tuned the bert-base-multilingual-cased model using HuggingFace Transformers[17]. Full sentences with token-level labels were used, with special handling of subword-token alignment and label masking. Training was performed for three epochs using a batch size of 8, 500 warm-up steps, and a weight decay of 0.01.

Additionally, we implemented a character-level BiLSTM model using PyTorch. Each token was encoded as a fixed-length sequence of up to 20 character indices from a shared Cyrillic vocabulary. The model architecture included a 64-dimensional character embedding layer, a bidirectional LSTM with 128 hidden units (64 per direction), a dropout layer (rate 0.5), and a final softmax output layer for binary classification. The model was trained from scratch for 10 epochs using the Adam optimizer (lr = 0.001) and the standard cross-entropy loss function. Importantly, like the other LSTM models in our experiments, this architecture operated on isolated tokens. However, by processing each token as a sequence of characters, the BiLSTM successfully captures sub-word patterns and morphological features, which are crucial for distinguishing between closely related languages like Ukrainian and Russian.

5 Evaluation and Analysis

5.1 Baseline Comparison and Evaluation

We evaluate our proposed methods by comparing them against several baseline systems, including GlotLID, a publicly available language identification library known for its strong performance on similar tasks (see Table 4). This comparison allows us to assess the effectiveness of our models relative to existing approaches.

Evaluation was conducted using standard metrics commonly employed in language identification tasks: Accuracy, Precision, Recall, and F_1-score. These metrics provide a comprehensive view of each model's overall correctness, relevance, and balance between false positives and false negatives. The comparative performance of all models is summarized in Table 5.

Among all models, BERT achieved the highest performance, with an F_1 score of 0.98, indicating its superior capability in token-level language identification. CRF with bigram features, The word-level neural classifiers (with and without pre-trained embeddings) and the character-level BiLSTM also performed well, achieving F1-scores around 0.90. This indicates that even without sentence-level context, lexical and morphological features are highly informative for this task. In contrast, the basic CRF model without n-gram features and the GlotLID baseline showed lower performance, highlighting the importance of contextual and sequential information in accurately identifying languages at the token level.

The implementation of our BERT-based token-level language identification model is publicly available on

[17] https://huggingface.co/google-bert/bert-base-multilingual-cased

Model	Accuracy	Precision	Recall	F_1
CRF (token features)	0.35	0.84	0.54	0.63
CRF (n-grams features, n=2)	0.93	0.99	0.95	0.97
Word-level NN	0.85	0.86	0.85	0.85
Word-level NN (pre-trained embeddings)	0.90	0.90	0.90	0.90
BERT	**0.98**	**0.98**	**0.98**	**0.98**
Char-level BiLSTM	0.90	0.90	0.90	0.90
GlotLID	0.76	0.82	0.77	0.75

Table 5: Performance metrics (Accuracy, Precision, Recall, and F_1 score) of all evaluated models on the test set.

Hugging Face[18], providing an easy-to-use resource for reproducing and building upon our results.

5.2 Qualitative Error Analysis

In this subsection, we discuss the error prediction analysis of BERT for language identification on the word level. We selected examples of both correct and incorrect predictions. The examples were chosen manually to illustrate both typical and challenging cases for the model. First, we examine examples of correct predictions. Russian text fragments are marked in bold.

(18) ***Вытолкали (ru)*** *цього (uk) провокатора (uk) , (punct) шановні (uk) друзі (uk) ! (punct)*

They pushed out *this provocateur, dear friends!*

For this example, the model worked correctly. The first word in this sentence was identified as Russian, but this is not a difficult case because the letter *ы* is only present in Russian. Therefore, this example can be considered a simple case for the model.

(19) *Я (uk) впевнена (uk) , (punct) що (uk) всі (uk) жінки (uk) проголосують (uk) , (punct) а (uk) чоловіки (uk) теж (uk), (punct) тому (uk) що (uk) пам'ятають (uk) слова (uk) великого (uk) письменника (uk) , (punct) що (uk) " (punct)* ***мерилом (ru) достоинства (ru) мужчины (ru) есть (ru) его (ru) отношение (ru) к (ru) женщине (ru)" (punct). (punct)***

*I'm sure that all women will vote, and men will too, because they remember the words of the great writer that "****a man's dignity is measured by his attitude toward women.****"*

(20) *Всі (uk) реєстри (uk) України (uk) можна (uk) купити (uk) на (uk)* ***книжном (ru) рынке (ru). (punct)***

All the registries of Ukraine can be bought at ***the book market.***

(21) ***Вы (ru) обязаны (ru) написать (ru) заявление (ru) с (ru) явкой (ru) повинной (ru)*** *і (uk) відповісти (uk) перед (uk) міліцією (uk)...*

You are obliged to write a confession statement *and answer to the police...*

Examples (19-21) are more complicated because they contain more words in Russian, and the words consist of letters that are present in both the Ukrainian and Russian alphabets.

Let us now examine an example of incorrect prediction.

(22) ***Уважаемый (ru)*** *головуючий (ru) ! (punct)*

Honourable *Chairperson!*

The sentence (22) contains only two words, one in Russian and one in Ukrainian, but the model identified both words as Russian.

(23) *Я (uk) звертаюсь (uk) до (uk) українців (uk) , (punct) кожен (uk) з (uk) нас (uk) буде (uk) заручником (uk) того (uk) , (punct) що (uk) згідно (uk) цих (uk) законів (uk) , (punct) які (uk) будуть (uk) прийняті (uk) , (punct) ви (uk) будете (uk)* ***обречены (ru)*** *на (ru) програш (ru) ...*

I am addressing Ukrainians, each of us will be a hostage to the fact that, according to these laws that will be adopted, you will be ***doomed to defeat...***

In (23), the word *обречены* 'doomed' was correctly identified as Russian, but the following part of the expression *на програш* 'to defeat', which is actually Ukrainian, was mistakenly classified as Russian by the model.

5.3 Comparison with GPT-3.5 model

We compared the performance of BERT for language identification with that of GPT-3.5 (December 2024). To conduct this comparison, we analyzed 100 sentences

[18] https://huggingface.co/OKanishcheva/ParlLangID-UA-RU

containing both Ukrainian and Russian words. These sentences were not included in the training corpus of the BERT model. We provide our prompts on our GitHub repository[19].

BERT and GPT-3.5 provided different answers in certain cases that are challenging to analyze. For example, this occurred when determining the language of prepositions shared by both languages (24). In our view, it is more appropriate to assign the language of the head noun to the preposition, given their close syntactic connection.

(24) BERT: *колега (uk) Купрейчик (uk) і (uk) наш (uk) колега (uk) із (uk) Дніпропетровська (uk) Безбах (uk) говорили (uk)* ***на (ru) русском (ru) языке (ru)***

GPT 3.5: *колега (uk) Купрейчик (uk) і (uk) наш (uk) колега (uk) із (uk) Дніпропетровська (uk) Безбах (uk) говорили (uk) на (uk)* ***русском (ru) языке (ru)***

Our colleague Kupreychik and our colleague from Dnipropetrovsk, Bezbakh, spoke ***in Russian.***

In some cases, the language of borrowed words is ambiguous. For instance, in example (25), the word мова *mova* 'language' is Ukrainian, but it appears in a Russian text as a barbarism to convey the speaker's negative attitude toward the Ukrainian language. In this context, the word is also found in other Russian texts and can thus be considered a borrowing. Similarly, in example (26), a Russian proper name Газпром *Gazprom* appearing within a Ukrainian sentence can likewise be interpreted as a borrowing.

(25) BERT: *керівник (uk) конвою (uk) сказав (uk) Надії (uk) Савченко (uk) : (punct) - (punct) "* ***(punct) почему (ru) это (ru) ты (ru) используешь (ru) эту (ru) чурбанскую (ru) мову (ru) " (punct)***

GPT 3.5: *керівник (uk) конвою (uk) сказав (uk) Надії (uk) Савченко (uk) : (punct) - (punct) "* ***(ru) почему (ru) это (ru) ты (ru) используешь (ru) эту (ru) чурбанскую (ru)*** *мову (uk) "* ***(punct)***

The convoy commander said to Nadiya Savchenko: ***"Why are you using that dumb language?"***

(26) BERT: *акт (uk) за (uk) січень (uk) цього (uk) року (uk) , (punct) між (uk) " (punct) Нафтогазом (uk) " (punct) та (uk) компанією (uk) " (uk) Газпром (uk) " (punct)*

GPT 3.5: *акт (uk) за (uk) січень (uk) цього (uk) року (uk) , (punct) між (uk) " (punct) Нафтогазом (uk) " (punct) та (uk) компанією (uk) " (punct)* ***Газпром (ru)*** *" (punct)*

The act for January of this year between "Naftogaz" and the company "Gazprom."

In example (27), BERT correctly identified the language of a Ukrainian word that was mistakenly written using the incorrect keyboard layout with Russian letters (*Включыть мікрофон* instead of *Включіть мікрофон*). GPT, however, failed to make this distinction.

(27) BERT: *Включыть (uk) мікрофон (uk)*

GPT 3.5: ***Включыть (ru)*** *мікрофон (uk)*

Turn on the microphone.

A total of 2,111 tokens were analyzed to compare the performance of the BERT model with GPT-3.5. In 81 cases, the models produced different predictions, which were manually reviewed. BERT gave correct answers in 58 of those cases, while GPT-3.5 gave correct answers in 4 of them and incorrect ones in the remaining 58. In the remaining 19 instances, the language of the token could not be determined unambiguously.

6 Conclusion

In this paper, we present a newly developed Ukrainian-Russian code-mixed dataset based on parliamentary transcripts from the Verkhovna Rada. The dataset was manually annotated at the token level to support the task of language identification. We detail the challenges encountered during annotation, particularly those arising from the linguistic similarity between Ukrainian and Russian.

We conducted a comprehensive analysis of the dataset, highlighting the complexity of distinguishing between closely related languages. To address the token-level language identification task, we evaluated several models, including CRF with various feature sets, LSTM, BiLSTM, and BERT. Our experimental results show that BERT achieves the best performance on this task. We also provide illustrative examples of the model's strengths and limitations in handling code-mixed language.

In the future, we plan to analyze the syntactic trees and discover the calquing of syntactic structures from one language into another. Additionally, we plan to explore a classification task for the *MIX* category (surzhyk); however, for this task, we need to collect more data. Currently, there already exist code-switching corpora annotated with morphological and syntactic information. Theoretically, adding these properties will allow for more accurate identification of the language, namely tokens from the *MIX* category, and enable research into the calquing of syntactic structures.

[19] https://github.com/olgakanishcheva/ukr-rus-code-switching-tagger

7 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'[20] (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.

References

Aguilar, Gustavo, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Aguilar, Gustavo and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Biletskyi-Nosenko, Pavlo. 1966. *Slovnyk ukrainskoi movy*. Nauk. dumka.

Bilodid, I. K., editor. 1970–1980. *Slovnyk ukrainskoi movy: v 11 tomakh [Dictionary of the Ukrainian Language: in 11 volumes]*. Naukova dumka, Kyiv. Academy of Sciences of the Ukrainian SSR, Institute of Linguistics named after O. O. Potebnia.

Burchell, Laurie, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Chavan, Tanmay, Omkar Gokhale, Aditya Kane, Shantanu Patankar, and Raviraj Joshi. 2023. My boli: Code-mixed Marathi-English corpora, pretrained language models and evaluation benchmarks. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 242–249, Nusa Dua, Bali. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erjavec, Tomaž, 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*.

Goswami, Koustava, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen, and John P. McCrae. 2020. Unsupervised deep language and dialect identification for short texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1606–1617, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Graves, A. and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

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

Hentschel, G. 2024. Ukrainian and Russian in the lexicon of Ukrainian Suržyk: reduced variation and stabilisation in central Ukraine and on the Black Sea coast. *Russ Linguist*, 48(2).

Hentschel, G. and O. Taranenko. 2022. Dvomovnist chy trykodovist: ukrainska mova, rosiiska mova i "surzhyk" v Ukraini (analiz i linhvistychnoheohrafichne kartohrafuvannia). *Movoznavstvo*, 1:21–50.

Hidayatullah, Ahmad Fathan, Rosyzie Anna Apong, Daphne T.C. Lai, and Atika Qazi. 2023. Corpus

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

creation and language identification for code-mixed indonesian-javanese-english tweets. *PeerJ. Computer science*, 9 e1312.

Jose, Navya, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Joshi, Ramchandra and Raviraj Joshi. 2020. Evaluating input representation for language identification in hindi-english code mixed text. *ArXiv*, abs/2011.11263.

Kanishcheva, Olha, 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.

Kargaran, Amir Hossein, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Khanuja, Simran, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

King, Ben, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mario, Kostelac. 2021. Comparison of language identification models.

Masenko, L. T. 1999. *Mova i polityka*. Soniashnyk.

Masenko, L. T. 2011. *Surzhyk: mizh movoiu i yazykom*. Kyievo-Mohylianska akademiia.

Mave, Deepthi, Suraj Maharjan, and Thamar Solorio. 2018a. Language identification and analysis of code-switched social media text. In *CodeSwitch@ACL*.

Mave, Deepthi, Suraj Maharjan, and Thamar Solorio. 2018b. 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.

Mikolchak, Maria. 2025. Codeswitching in Ukrainian media: Social meaning of Ukrainian-Russian bilingualism in Ukraine. *Journal of Arts and Humanities*, 14(3):51—61. Published 29 June 2025 (Vol.14, No.03).

Mozer, M. 2016. "Surzhyk" chy "surzhyky"? ["Surzhyk" or "surzhyks"?]. *Ukrainska mova*, (1):27–54.

Pylypenko, Valeriy and Tetyana Lyudovyk. 2019. Automatic recognition of mixed Ukrainian-Russian speech.

Rani, Priya, John P. Mccrae, and Theodorus Fransen. 2022. Mhe: Code-mixed corpora for similar language identification. In *International Conference on Language Resources and Evaluation*.

Schmidt, Anastasia. 2014. *Between The Languages: Code-Switching in Bilingual Communication*. Anchor Academic Publishing, Hamburg.

Shvedova, M. 2016. Inojazyčnye vlijanija v russkoj reči Ukrainy (Leksika) [Foreign-language influences in the Russian speech of Ukraine (Lexicon)]. *Movoznavčyj visnyk: Zbirnyk naukovyx prac'*, 21:86–92.

Shvedova, Maria, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, Tymofij Nikolajenko, et al. 2017–2025. GRAC: General regionally annotated corpus of Ukrainian. Electronic resource. Available at https://uacorpus.org.

Sira, Nataliya, Giorgio Maria Di Nunzio, and Viviana Nosilia. 2019. Towards an automatic recognition of mixed languages: The Ukrainian-Russian hybrid language Surzhyk.

Taranenko, Oleksandr. 2007. Ukrainian and Russian in contact: attraction and estrangement. *International Journal of the Sociology of Language*, 2007(183):119–140.

Thara, S. and Prabaharan Poornachandran. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837–118850.

Winata, Genta, 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.

Zhang, Yuan, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.