Creation Dataset of Token Language Identification for Ukrainian-Russian Code-switching Corpus

Olha Kanishcheva
University of Jena,
SET University
kanichshevaolga@gmail.com

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

Relevant UniDive working groups: WG1

1 Introduction

In the dynamic landscape of natural language processing (NLP), the analysis of code-switching, where speakers seamlessly blend multiple languages within a single discourse, poses a unique set of challenges (Barik et al., 2019). The Ukrainian linguistic context, rich in cultural diversity and historical influences, is a fascinating domain for exploring code-switching phenomena. This article delves into the intricate task of language identification within Ukrainian code-switching corpora, shedding light on the complexities inherent in deciphering linguistic boundaries in a multilingual environment.

The amalgamation of languages in code-switching scenarios not only reflects the sociolinguistic intricacies of a community but also presents a captivating puzzle for NLP researchers. Within the Ukrainian context, where bilingualism and multilingualism are pervasive, understanding and accurately identifying the languages involved in code-switched utterances become paramount for developing robust language processing systems.

This article discusses ongoing, unfinished research. It aims to explore the peculiarities of code-switching in Ukrainian corpora, highlighting the complexities in processing texts where Ukrainian and Russian are present and Ukrainian-Russian mixed speech (Surzhik), which contains hybridization within a word. The work will also outline an approach to identifying languages, illustrated by the Code-Switch Parliamentary Corpus as an example.

2 Features of Ukrainian Corpora and Data

Language detection for the purpose of extracting or tagging foreign language fragments is an important part of processing Ukrainian corpora of various types, although it is primarily concerned with less standardized texts such as spoken language or Internet communication. The modern Ukrainian language was standardized in the early twentieth century. During most of its history, it developed in conditions of bilingualism, with Polish (in western Ukraine until 1939) or Russian (until independence in 1991) as the dominant language. In the decades since Ukraine became independent, the Ukrainian language has been actively developed, being used in various social spheres, and the number of speakers has been increasing. At the same time, a large part of the population of modern Ukraine is bilingual and also uses Russian as a spoken language in everyday life. A significant part of Ukrainian spoken data includes also elements of dialects, Ukrainian colloquial speech, and Ukrainian-Russian mixed speech, which is highly variable and therefore difficult to describe formally (Mozer, 2016).

Thus, when processing Ukrainian language data from different historical periods, we face various challenges caused by the history of bilingualism.

Until the 1990s, Ukrainian texts contained fragments in Russian, most often quotes without translation. In older works of Ukrainian fiction, Russian in the speech of characters is most often presented without translation. The same is true for Polish in old Western Ukrainian literature, but Polish inserts are less of a problem for processing Ukrainian texts, since Polish uses the Latin alphabet, and Russian, like Ukrainian, uses the Cyrillic alphabet, which differs only in a few letters.

In the late 1990s, Ukrainian language data downloaded from the Internet became available. Later, automatic translation became widespread, and for such closely related languages as Ukrainian and Russian, it worked well enough. Many websites had parallel Ukrainian and Russian versions until 2014 and even some until 2022, where Ukrainian was an automatic translation from Russian. Determining the language when downloading texts from the Internet is not a problem as such, but determining where the original text is and where the result of automatic translation is more difficult. Therefore, the first web corpora of the Ukrainian language contain a large share of low-quality data.

When it comes to modern data, the biggest problem for processing language switching and mixing is non-standard texts: spoken recordings and Internet communication. This type of data is increasingly found in the modern corpora, e.g. GRAC (Shvedova et al., 2017-2023), in transcripts of spoken language and unnormalized texts from the Internet. Processing of such texts using systems developed for a standard language is not always successful. Currently, the identification of code-switching in the GRAC corpus is performed using the approach described in (Starko et al., 2021). The approach is not perfect because it does not annotate the language at the token level, so there is a need to create new tools.

3 Corpus Creation and Annotation

In recent decades, more and more studies have explored token-level language identification for different languages. In works (Winata et al., 2023; Hidayatullah et al., 2022) the authors presented the analysis of different tasks related to code-switching corpora, different corpora, issues with processing of such data, etc. In our research, we specifically targeted sentences within parliamentary transcripts of the Verkhovna Rada that exhibited a fusion of Ukrainian and Russian languages. This approach yielded a dataset comprising approximately 150,000 tokens.

The dataset consists of separate sentences selected from the corpus of Ukrainian parliamentary transcripts (Kanishcheva et al., 2023). We excluded from the corpus sentences in Russian. After that, we lemmatized the corpus using the Ukrainian dictionary[1] and selected sentences with more than two unknown words. In the vast majority of cases, these were sentences with some words in Ukrainian and some in Russian. A small number of sentences contained words with errors or non-dictionary words.

All sentences were tokenized and each token was labeled. The labels are used summarized in Table 1.

Labels	**Description**	**Tokens**
UK	Ukrainian words	93 040
RU	Russian words	30 956
Mix	Ukrainian-Russian hybridised words (Surzhyk)	225
Others	Dialects, other languages, etc.	615
Punct	Punctuation	30 695

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

Categorization involved the allocation of tokens into five distinct classes: Ukrainian, Russian, Ukrainian-Russian hybridised words (Surzhyk), Others, and Punctuation.

Examples of language annotation in the text: *ce <uk> tjahne <uk> za <uk> soboju <uk> rist <uk> ciny <uk> na <uk> spirtosoderžaščie <ru> lekarstva <ru> , jak <uk> to <uk> ukraïns’koju <uk> movoju <uk> skazaty <uk> ?* (This entails an increase in the price of *alcohol-containing drugs*, how to say it in Ukrainian?); ja <uk> robotav <mix> , včyvsja <uk> v <uk> Xarkovi <uk> (I *worked*, studied in Kharkiv).

In many cases, the unambiguous distribution of tokens into these categories proved problematic. The data contains many words that spell the same in Ukrainian and Russian. They are marked as Ukrainian or Russian depending on the context, according to the language of the syntagm. However, there are often cases when such words are on the borderline between Ukrainian and Russian. Attribution of such a word as Ukrainian, Russian, or hybridised is impossible without listening to a recording of the pronunciation, for example: *prošu <ru> ... postavit’ <ru> ėtu <ru> popravku <ru> Burjaka <ru> na <ru> golosovanie <ru> i <uk|ru??> prošu <uk|ru??> zal <uk|ru??> ne <uk> pidtrymuvat’ <uk> cju <uk> popravku <uk>* (’I ask you to put this amendment by Buriak to the vote *and ask the chamber* not to support this amendment’).

In some cases, the stenographer did not follow the spelling standard; such cases were corrected in the data, but often proved to be helpful as they indicated the speaker’s pronunciation. A complicated case is expressions translated from Russian, where each word is Ukrainian and the whole phrase is hybridised. In such cases, it is impossible to define language only at the word level, for exam-

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

ple, the phrase: *A ščo torkajet'sja zakonoproektu...* (uk: *'As for the draft law...'*) is a calque from Russian *A čto kasaetsja zakonoproekta...*, as Ukrainian *torkatysja* 'to touch', unlike Russian *kasat'sja*, does not have the figurative meaning 'to relate to'.

4 Overview of Language Identification Libraries

Language identification is a natural language processing task aimed at determining the language of a given piece of text. This task holds significant importance in a variety of applications, ranging from information retrieval and machine translation to sentiment analysis and content filtering. Challenges in language identification include dealing with short texts that may lack sufficient linguistic patterns and distinguishing between languages that share similarities, such as Spanish and Portuguese or Ukrainian and Russian.

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,[2] Fasttext,[3] langid.py,[4] Spacy,[5] CLD3,[6] Langdetect,[7] and Lingua[8]. These tools provide efficient ways to implement language identification algorithms, making it accessible for developers and researchers alike.

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

5 Description of General Approach

Several tasks are planned for our study. The first one is to evaluate the accuracy of language identification by different libraries (discussed in Section 2) at the token level using an annotated dataset. Evaluations of the performance of different libraries for identifying close languages such as Ukrainian and Russian have already been performed, but not for code-switching data. This study will be conducted for the first time.

The next task of this research is to build a classification model that will be able to determine the language at the token level for such classes as Ukrainian, Russian, Surzhik, and others.

We plan to apply methods to the task of token classification such as Support Vector Machine (SVM), Conditional Random Fields (CRF) (with N-grams on different levels), and some Transformer methods (BERT, ELECTRA, etc.) (Chavan et al., 2023; Doğruöz et al., 2021). These methods will evaluate how the proposed methods cope with the task of inter-word code-switching identification for Slavic languages.

6 Conclusions and Future Work

At this stage of work, a dataset of about 150,000 tokens has been collected, which contains code-switching between Ukrainian and Russian languages. Also, this dataset contains intra-word code-mixing, so-called Surzhik. This dataset is divided at the token level into 5 categories. The next stage will be to analyze the obtained dataset and test different classification models on this data.

The development of language detection tools is important to improve the annotation of existing Ukrainian corpora and the creation of future ones, as Russian infiltrations and mixing are frequent problems in Ukrainian data. Particularly, there are plans to use the dataset for language annotation at the token level in Ukrainian ParlaMint.[9]

Acknowledgements

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

This research was partially funded by the Humboldt Foundation and the Volkswagen Foundation.

References

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English

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

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

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

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

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

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

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

[9] https://www.clarin.si/repository/xmlui/handle/11356/1900

code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

Laurie Burchell, 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.

Tanmay Chavan, Omkar Gokhale, Aditya Kane, Shantanu Patankar, and Raviraj Joshi. 2023. My boli: Code-mixed marathi-english corpora, pretrained language models and evaluation benchmarks. *CoRR*, abs/2306.14030.

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.

Koustava Goswami, 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.

Ahmad Fathan Hidayatullah, Atika Qazi, Daphne Teck Ching Lai, and Rosyzie Anna Apong. 2022. A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development. *IEEE Access*, 10:122812–122831.

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.

Kostelac Mario. 2021. Comparison of language identification models.

M. Mozer. 2016. "suržyk" čy "suržyky"? *Ukraïns'ka mova*, 1:27–54.

Maria Shvedova, Ruprecht von Waldenfels, Sergey Yarygin, Andriy Rysin, Vasyl Starko, and Tymofij Nikolajenko et al. 2017-2023. *GRAC: General Regionally Annotated Corpus of Ukrainian*. Electronic resource: Kyiv, Lviv, Jena, Available at uacorpus.org.

Vasyl Starko, Andriy Rysin, and Maria Shvedova. 2021. Ukrainian text preprocessing in grac. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, volume 2, pages 101–104.

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.