WG1

UniDive
2nd General Meeting
University of Naples "L'Orientale
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

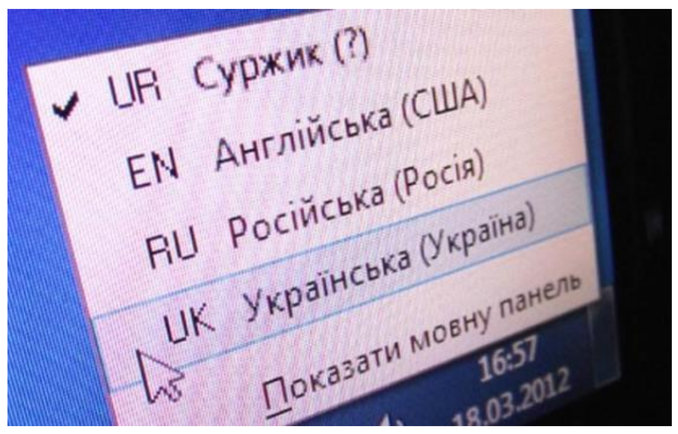
Funded by the European Union

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

Olha Kanishcheva[1,2], Maria Shvedova[1,3]

[1]University of Jena, [2] SET University, [3] National Technical University "KhPI"

Introduction



https://www.ukrinform.ua/rubric-society/3408444-surzik-i-cistota-abo-comu-ce-vidvratitelno.html

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.

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.

Types Of Code-Switching

- **Intersentential**: Intersentential code-switching occurs outside the sentence or clause. In other words, a complete sentence/clause in one language is followed by one in another language. It is also called "extrasentential switching".
- **Intrasentential**: Intersentential code-switching occurs within the sentence or clause. A part of the sentence is in one language/language variety and is then followed by one from another language/language variety.
- **Tag-level**: Tag-level switching brings a tag phrase or a word from one language into another language.
- **Intra-word**: Intra-word switching occurs within a word, such as at the morpheme level.

Corpus Statistics

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

The dataset consists of separate sentences selected from the corpus of Ukrainian parliamentary transcripts[1]. We excluded from the corpus sentences in Russian. After that, we lemmatized the corpus using the Ukrainian dictionary[2] 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.

[1]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.

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

Examples

2) Pry (uk) c'omu (uk) same (uk) Rosija (uk) , a (uk) ne (uk) Ukraïna (uk) provodyt' (uk) zasidannja (uk) Radbezu (uk) z (uk) pytannja (uk) ščodo (uk) ukrainskogo (ru) cerkovnogo (ru) voprosa (ru).

(At the same time, it is Russia, not Ukraine, that is holding the Security Council meeting on the Ukrainian church issue.)

3) Zakonoproekt (uk) pro (uk) zaboronu (uk) propahandy (uk) ideolohiï (uk) ' russkogo (ru) mira (ru)' je (uk) očikuvanym (uk) ta (uk) aktual'nym (uk)

(The bill banning the propaganda of the "Russian world" ideology is expected and relevant)

4) Naviščo (uk) tut (uk) portyty (mix) nervnu (mix) systemu (uk)

(Why injure the nervous system here?)

Moju (uk) komandu (uk) prynyžuje (uk) ocja (uk) vaša (uk) myšyna (mix) voznja (mix) za (uk) komitety (uk)

(My team is humiliated by your mouse fussing over committees.)

Code-switching analysis

Metrics

- Lexical Overlap
- Code-mixing Index

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

- M-Index

$$M - Index = \frac{1 - \sum p_j^2}{(k-1)\sum p_j^2}$$

- Integration Index

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

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.

Code-Mixing Index (CMI) : An utterance level metric that measures the fraction of tokens(words) that are not from the matrix language.

$$CMI = 100 * (1 - \frac{max(w_i)}{n-u}) \text{ if } n > u$$

where,

n = total number of words, irrespective of language

u = language independent words

w_i = number of words in language i

$max(w_i)$ measures the number of words in the matrix language

If $n = u$ i.e. if the utterance contains only language independent words then CMI = 0.

Language Pair	CMI
English-Bengali (Gambäck and Das, 2014)	24.48
Dutch-Turkish (Nguyen and Doğruöz, 2013)	22.65
Modern Arabic-Egyptian Arabic (Molina et al., 2016)	3.89
Spanish-English (Mave et al., 2018)	22.11
Hindi-English (Mave et al., 2018)	22.22
Nepali-English (Solorio et al., 2014)	20.32
Magahi-Hindi-English	51.54

Code-Mixing Index for our data: 37.04%

One way to calculate CMI for the entire corpus is to just calculate the above at a corpus level rather than an utterance level. However, this method doesn't take into account the switching frequency.

Conclusions

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.

https://www.clarin.eu/parlamint

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

- Olha Kanishcheva kanichshevaolga@gmail.com
- Maria Shvedova mariia.shvedova@khpi.edu.ua