Umanistica Digitale – ISSN: 2532-8816 – n. 9, 2020
N. Sira, G. M. di Nunzio, V. Nosilia – *Towards an Automatic Recognition of Mixed Languages: The Case of Ukrainian-Russian Hybrid Language Surzhyk*
DOI: http://doi.org/10.6092/issn.2532-8816/10740

Towards an Automatic Recognition of Mixed Languages: The Case of Ukrainian-Russian Hybrid Language Surzhyk

[1]Nataliya Sira, [2]Giorgio Maria di Nunzio, [3]Viviana Nosilia

[1]RWTH Aachen University, Germany
[2,3]University of Padua, Italy
[1]nsira@ukaachen.de
[2]giorgiomaria.dinunzio@unipd.it
[3]viviana.nosilia@unipd.it

Abstract

Language interference is common in today's multilingual societies where more languages are in contact. As a global result, it leads to the creation of hybrid languages. These, together with doubts on their right to be officially recognised, highlight the problem of their automatic identification and further elaboration in the area of computational linguistics. In this paper, we propose a first attempt to identify the elements of a Ukrainian-Russian hybrid language, Surzhyk, through the adoption of the example-based rules created with the instruments of programming language R. Our example-based study consists of: 1) analysis of spoken samples of Surzhyk registered by Del Gaudio ([15]) in Kyiv area and creation of the written corpus; 2) production of specific rules on the identification of Surzhyk patterns and their implementation; 3) testing the code and analysing the effectiveness of the hybrid language classifier.[1]

Il concetto di interferenza linguistica è diventato caratterizzante nelle società multilingue odierne, nelle quali più e più lingue entrano in contatto e portano alla creazione di lingue ibride. Tali lingue, assieme alle discussioni sul loro diritto di essere riconosciute a livello ufficiale, hanno fatto emergere il problema della loro identificazione e elaborazione automatica nell'area della linguistica computazionale. Nel presente articolo proponiamo un primo tentativo di identificare gli elementi di una lingua ibrida ucraino-russa, il suržyk, attraverso l'adozione di regole basate sugli esempi create con gli strumenti del linguaggio di programmazione R. Il nostro studio basato sugli esempi consiste in: 1) analisi delle registrazioni di suržyk parlato raccolte da Del Gaudio ([15]) nell'area di Kyjiv e creazione di un corpus scritto; 2) produzione e implementazione di

1 Some parts of the work for this research were developed in the context of Nataliya Sira's Master thesis project at the Department of Linguistics and Literary Studies, University of Padua, Italy.

regole specifiche sull'identificazione degli elementi di suržyk; 3) test del codice creato e analisi di efficacia del classificatore della lingua ibrida.

Introduction

In the multilingual language use, the phenomenon of code mixing has become common, and it might be considered a natural product of multilingualism. In social media communication, for example, multilingual speakers often switch between languages ([11]). The former is just one of the proofs that confirms a rapid and irreversible development of mixed languages and, as a consequence, the problem of their identification. According to Barman and others ([11]), in those situations where speakers switch between languages or mix them, the automatic language identification process is increasingly important as it facilitates further language processing.

Among the modern mixed languages, Surzhyk is a hybrid language that involves Ukrainian and Russian languages in its creation. Similarly to other newly emerged mixed languages, Surzhyk has the problem of digital data scarcity; this means not only that scientific studies on Surzhyk structure are limited but also that there is no digital corpus of Surzhyk ready to be used. Although Surzhyk prestige is remarkably low both from the point of view of experts and that of the common public ([22]), this spoken language is commonly used in the whole of Ukraine. Even though we do not have a reliable estimation on the number of Surzhyk speakers, we can identify that the largest number of Surzhyk speakers was registered in eastern, southern and central parts of Ukraine ([19]). Moreover, it should be considered that in some cases its structure and lexicon may be subject to the influence of the dialects present in the Ukrainian territory, and it can vary depending on the speakers' characteristics.

Given this particular condition, Surzhyk is an interesting case study for automatic language identification since it is at the same time relatively widely spoken and scarce of linguistic resources.

The reminder of this paper is organised as follows. In the introductory part, we present an overview of the background studies and the objective of this study. Section 2 contains the preliminary analysis on the definition of patterns in Surzhyk. In Section 3, we describe the process of data collection and elaboration; Section 4 presents the rules for automatic identification of Surzhyk patterns; in Section 5, we present the experimental setting and analysis of results. Finally, Section 6 is a discussion followed by Section 7 with conclusions and proposals on possible developments of the current study.

The Concept of Hybridity and Mixed Languages

Language interference is a common topic in today's society. In the areas of official or unofficial multilingualism, it is inevitable that languages come in contact and somehow influence each other. In the situations of language contact, a completely new language may emerge that would present elements of both languages involved in its creation ([6]). The hybridity of the mixed languages consists of the fact that they take their lexicon from one source and grammar from

another. Their classification becomes difficult as, on the basis of the lexicon, they could appartain to one language family, and on the basis of morphology, syntax, and general grammatical characteristics they may belong to another language family ([10]). According to the criterion aimed at defining the nature of mixed language provided by Bakker ([8]), "a language is mixed if it can with equal justification be assigned to two different language families."

Ukrainian-Russian Surzhyk as an Example of Hybrid Language with Mixed Language Characteristics

Following the argument of Bakker on the case of Michif, the Cree-French mixed language of the Métis, where "both French and Cree speakers may say it is their language when they recognise roughly half of what is said, but they have to admit that much of it remains unintelligible", one can see some similarities in the phenomena of Surzhyk. The criterion proposed by Bakker may be applied to the cases where "the lexicon is from one language and the grammatical system (phonology, morphology, and syntax) from another" ([8]). In Surzhyk, this seems to be the case, especially if we consider that the process of "intertwining" mixes lexicon and grammar of different codes and leads to the creation of the new language that typically presents stems from one language and affixes from another ([9]). In the current research, the idea of the Matrix Language Frame model (MLF) elaborated by Myers-Scotton ([23],[24]) and proposed by Kent ([19]) for Surzhyk analysis is taken into consideration. According to the MLF model, only one language is the source of the abstract morphosyntactic frame in a bilingual clause, and that is the Matrix Language (ML). The other participating language is the Embedded Language (EL), and it must agree with structural requirements stipulated by the ML ([23],[24]). According to Kent ([19]), in Ukrainian-Russian Surzhyk structure, Ukrainian is a ML whereas Russian is an EL. The nature of Surzhyk is complex: it has a huge variety of possible mutations based on the geographical area and speaker's characteristics (age, dominant language, type of education, social affiliation, and language experience).

According to Olszański, it could be admitted that Surzhyk is a Ukraine-specific language phenomenon[2] enabled by the co-existence of two mutually intelligible languages[3] in the same territory ([25]). The presence of both Ukrainian and Russian languages in the press, television, and radio should guarantee good knowledge of Ukrainian and Russian. However, what happens in practice is that these two languages interact and interfere with each other, which often gives space to hybrid speech.

2 Trasianka, present in Belarus, has a similar nature as Surzhyk.

3 We considered different thoughts in this topic but the priority of our study was concentrated on the possibility to analyse Surzhyk automatically rather than discussing whether Russian and Ukrainian may be classified as mutually intelligible languages or not. However, we prefer to present to the reader that there are different points of view. Relying on the Lewis' classification of languages' mutual intelligibility, Russian and Ukrainian result to be mutually unintelligible languages though [20].

NLP and Text Mining for Language Recognition

Natural Language Processing (NLP) is an area of computer science that researches how computer systems can analyse, understand, or produce natural languages. All possible human languages are natural languages, and they can be available as text, spoken language, or keyboard input ([4]). NLP was founded on the basis of a number of different research areas, such as computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence, and psychology ([14]).

NLP has a variety of subfields. An important area of research in NLP is natural language text processing. It enables the elaboration and structuring of large bodies of textual information with the purpose to retrieve particular information ([14]). One of the NLP subfields that closely deals with this research is text mining. Text mining is an interdisciplinary field that includes different theoretical approaches and methods that deal with the same type of input information—text. Text mining analysis includes the import of the text into a programming environment; the organisation of the text structure; text pre-processing that often consists in whitespace removal, stopword elimination, and stemming process; and the creation of the term-document matrix that is a common format used for text representation in computation. This enables the representation of a text, which is inherently unstructured to a computer, in a highly well-structured manner. A text mining framework offers a range of functionalities aimed at managing text documents and at allowing efficient work with them. Moreover, it provides filtering functionalities that may extract patterns of interest from text collections ([16]).

The issue of how to mine this type of information from large text collections in an efficient and effective way has been studied by means of organised workflows named pipelines ([35]).

Pipelines are an effective way to manage the sequential process of text analysis by splitting the source code into steps, where the output of one step is the input for the subsequent step. Apart from being a tidy way of organising software,[4] an important advantage in working with pipelines is that this practice promotes shareability and reproducibility in research workflows.

In this work, we use the tidyverse approach of the R programming language, which follows the principles of "pipelines", in order to design and implement the software for the analysis and identification of Surzhyk.

Objective of the Study

Since the development of mixed languages is rapid and irreversible, the study aims to demonstrate first attempts on the automatic identification of such languages. In our case, we propose an example-based study of Surzhyk samples recorded in different areas of Ukraine and their elaboration in the R programming language. This study aims at identifying particular patterns of Surzhyk verbs. The adoption of an example-based method was defined by the fact that Surzhyk is a less resourced language with a limited corpus and no complete grammar description available. Thus, we decided to create a digital text corpus of Surzhyk corresponding

4 https://www.tidyverse.org

to the recordings of real interviews collected by Del Gaudio ([15]). After the creation of the text corpus of Surzhyk, we studied the terminology present in this corpus and produced the terminological tables that we used to identify the patterns of the Surzhyk language. After a careful selection of the patterns, we implemented the classification rules, and we tested the effectiveness of this classification. The corpus as well as the source code is available on Github for reproducibility purposes.[5]

Preliminary Analysis

In this section, we present background studies on Surzhyk patterns definition. The decision to analyse two particular characteristics of Surzhyk verbs was taken with regard to the recurring repetition of these in the processed texts that was noticed during the process of manual transcription of the registrations collected by Del Gaudio ([15]). Moreover, it is known that morphology is a good marker for identification of language use. Firstly, we focused on the first-person plural Surzhyk ending of the present tense "-м". This decision was based on the fact that the ending "-м" resulted to be a recurring marker of Surzhyk verbs in the first-person plural. Secondly, we analysed the prefix "под-" of Surzhyk verbs, which, based on our samples, resulted to be another evident marker of Surzhyk verbs. To carry out this analysis, it was necessary to compare the Russian and Ukrainian verbal systems, and, specifically, the ending of the first-person plural in the present tense. Together with the facts on historic development of the first-person plural endings in present tense in Slavonic languages, the current differences of the endings in both Russian and Ukrainian languages were taken into consideration. Regarding the analysis of the second particle of interest, we took into consideration the word formation system in Russian and in Ukrainian and compared prefixes that are usually adopted in the prefixation process.

The Study of the First-Person Plural Ending Form "-м"

In Ukrainian and in Russian there are two verb conjugations: conjugation 1 and conjugation 2.

Russian verbs belonging to the conjugation 1 are characterised by the following endings in present tense: **-ю (-у), -ешь (ёшь), -ет (-ёт), -ем (-ём), -ете (-ёте), -ют (-ут)**. Russian verbs of the conjugation 2 are characterised by the following endings in present tense: **-ю (-у), -ишь, -ит, -им, -ите, -ят (-ат)** ([3]). In Ukrainian, the endings of the conjugation 1 in present tense are: **-у (-ю)**, **-еш (-єш)**, **-е (-є)**, **-емо (-ємо)**, **-ете (-єте)**, **-уть (-ють)** ([32]). Ukrainian verbs of the conjugation 2 are characterised by the following endings in present tense: **-у (-ю)**, **-иш (-їш)**, **-ить (-їть)**, **-имо (-їмо)**, **-ите (-їте)**, **-ать (-ять)**.

5 https://github.com/gmdn/Surzhyk

In the first-person plural of the present tense, a vowel that we have considered as a part of the ending prior can—from another point of view—be seen as a thematic or connector vowel between a stem and an ending of a verb. Consequently, we can admit that possible endings of the present tense in first-person plural can be **-мо** in standard Ukrainian or **-м** in standard Russian. The difference that nowadays seems to be clear in two standards was not so definite before, especially for the Ukrainian language. A Ukrainian grammar of 1927 presents the ending **-м** as possible variant to the ending **-мо** in the first-person plural ([13]).

"The ending **-мо** has been preserved in Ukrainian till nowadays" ([21]). Though the inflection **-мо** is used more often, the ending **-м** can also appear, even within the text or speech of the same source where the form **-мо** is used. The first form is fixed as the leader within grammatical and lexicographical literature, while within the language of writers from both Eastern and Western Ukraine, as well as in the language of folklore, both forms can coexist ([1]).

The Study of the Prefix "под-"

In the Russian language, the prefix **под-** partially corresponds to the preposition **под** (which means "under") in its meaning. The adoption of this prefix adds the following meaning to the verbs: the direction of action goes under the subject; the movement is from bottom to top; approximation; incompleteness of the actions when a) the action does not cover the whole object, but only its surface part or/and b) the action is limited by time, performed only on this occasion, for this purpose, or slightly; the action is carried out secretly. The accompanying action is expressed by the prefix **под-** and by the suffix **-ыва-** (**-ива-**)—typical markers for the imperfective aspect ([2]).

In modern Ukrainian language, there is a large group of verbs that were formed long ago; such verbs form the basis for the creation of new, derived verbs. There are three ways of forming verbs: by adding prefix, suffix, or both suffix and prefix. The first method of verb formation—by adding a prefix—is common for new verbs deriving from the verbal bases (also called internal verb word formation). This is verbal word-formation ([12])[6] . Modern Ukrainian language does not present a "под-" prefix; a regular Ukrainian prefix corresponding to the prefix **под-** we analysed in the Surzhyk samples would be a prefix **під-**. It should be noted that in the Ukrainian lexicon there are words with a **по-** prefix and—when followed by a consonant **д**—at first glance they may deceive a non-specialist of the area, especially when the morphology of the word is not taken into consideration.

Thus, the prefix **под-** reveals itself to be Russian, but in combination with Ukrainian verbs it becomes an important Surzhyk marker. The meaning of the Surzhyk prefix **под-** corresponds to the Russian prefix **под-** and Ukrainian **під**-. All in all, it seems to be a significant and characteristic particle in the identification of Surzhyk verbs.

6 Please check this reference for more information on the most important prefixes involved in the derivation process of Ukrainian verbs.

Data Collection

Surzhyk is more characteristic as a spoken language rather than written, it has no structured corpus nor normalisation. Moreover written samples of spontaneously spoken Surzhyk are not available, thus the collection of Surzhyk samples was a critical task. During the first stage of the research we analysed the interviews recorded by Del Gaudio in different areas of Ukraine that were published as a CD together with his PhD dissertation *On the nature of Suržyk* ([15]). The interviews of Del Gaudio involved people from Kyiv, Chernihiv, and Kharkiv areas, but in the current research only the first group of conversations regarding the area of Kyiv were elaborated on in detail. In order to have digital data, it was necessary to manually transcribe all of the recordings since no effective software for speech-to-text automatic recognition was available for Ukrainian, nor for the non-standard spoken language Surzhyk. The first stage of the analysis consisted in creating the Surzhyk corpus. In order to have the material in the proper format for computation and further elaboration, all the recordings collected by Del Gaudio were transcribed manually and saved as digital text documents. The process was guided by the rule of writing what was heard, and Ukrainian writing rules were utilised as the basis. It is important to note that audio registrations were not always noiseless and clear. For these reasons, there may be some omissions in the corpus. In addition, there may be some small typing errors. Secondly, in the previously created text documents we identified lexical elements that did not belong to the standard Ukrainian language, as well as words that were defined as incorrect Ukrainian by an automatic spell checker. This analysis was limited and mainly consisted of identifying the non-standard Ukrainian lexical elements in the text documents and creating their representation in spreadsheet documents. Consequently, we present the structure adopted in these documents.

Surzhyk term	Ukrainian written	Russian written	Russian pronounced	Additional comments	Context	English written
каждого	кожного	каждого	каждава	*(none)*	в каждого	each (in genitive)
покупає	купляє	покупает	пакупаєт	*(none)*	*(none)*	to buy (3rd pers. sing.)

Table 1: The structure of the simplified terminological tables of Surzhyk.

In order to analyse every term in a proper way, we compare every word to the terms in Ukrainian and in Russian, and we also provide an English translation. While for Ukrainian there is only the corresponding written version, for the Russian terms we both present the written version that respects the rules of Russian and the phonetic pronunciation of the term according to Ukrainian phonology. This becomes necessary since "Russian content morphemes introduced to the Ukrainian morphosyntax are subject to Ukrainian phonology and are often pronounced in accordance with Ukrainian phonological rules" ([19]). In this way, we created and filled the terminological records that in total present more than 1000 non-standard Ukrainian terms. Some may argue that a term that corresponds perfectly to the Russian pronunciation should be

classified as a Russian one, but we always have to consider the following facts: first, the identified elements are characteristic of and widespread as a non-standard spoken language; second, in this type of analysis we elaborate texts that were registered and collected in Ukraine and the main language of the conversation was Ukrainian. If we consider the MLF model and the fact that in the analysed texts we have Ukrainian as a ML and Russian as an EL, our samples should contain a great amount of Russian lexical items and mostly Ukrainian morphosyntax. Therefore, if the main language of the conversation was Russian, we would surely affirm that terms that correspond to the Russian pronunciation are Russian, but in this specific analysis we are trying to define the spoken language, that some speakers define as "not a clean language" ([15]), a some kind of "mixed Ukrainian" ([19]), "this type of Ukrainian" ([19]). During this research phase, we produced simplified terminological tables of Surzhyk. Simplified terminological tables of Surzhyk are composed of seven general tables presenting all the hybrid elements we found in the analysed texts and correspond to seven records of the Kyiv area collected by Del Gaudio. We maintained the original titles of the audio files when giving names to our transcribed versions, but the references to the files in this paper contain only the shortened titles.[7]

Additionally, we elaborated on a model of the deeper description of the hybrid language terms (standard terminological table of Surzhyk). The structure of the standard terminological table of the Surzhyk model was drafted in order to define important aspects of Surzhyk terms. Although creating terminological tables for Surzhyk was only part of the preparatory phase and not the main task of this research, we consider that they may be useful for further studies on Surzhyk.

Rules for the Automatic Identification of Surzhyk Patterns

In this section, we present the linguistic patterns that we defined in order to automatically identify the Surzhyk hybrid language.

Final Pattern "-м" Characterising the First-Person Plural Verbs in Present Tense

Two groups of rules were created: general and specific. Subsequently, we present the rules of both groups in the descriptive form.[8]

1st pattern (patt.)	2nd patt.	Distance	Output word requests
ми	-м	max. 3	1st patt. = "ми"
самі	-м	max. 3	
-м	-ти	max. 3	2nd patt. > "ти"

7 For example, instead of using the full title "DW_A0039.docx" that refers to the audio file and to the transcribed version of the audio, we adopted the shortened title "39".

8 In this paper "present" is used conventionally due to the space limits, but it always intends to refer to "present / future," depending on the verb aspect.

-м	-ть	max. 3	

Table 2: Surzhyk general rules on identification of the first-person plural verb in present tense.

The first two rules determine that the first element is a word "ми" or "самі" while the second element "-м" is a final part of the word. The second element follows the first and is situated on a maximum distance of 3 words from the first element.

The second two rules imply that the first element ends in "-м" while the second one ends in "-ти" or "-ть" and is situated on a maximum distance of 3 from the first element.

The general rules also present requirements which have to be respected by the output word. In the case of the first general rule, the first pattern of the output has to be a single word "ми" and not a part of a word. A similar requirement exists within the second pattern of the third rule: the second pattern of the output that ends in "-ти" has to be a word with a number of characters higher than the number of characters present in the word "ти" since it would be a personal pronoun of the second-person singular and not a verb ending.

1st patt.	2nd patt.	Distance	Output word requests
ми	-ем	max. 3	
ми	-єм	max. 3	
ми	-им	max. 3	
ми	-їм	max. 3	2nd patt. > "їм"
самі	-ем	max. 3	
самі	-єм	max. 3	
самі	-им	max. 3	
самі	-їм	max. 3	2nd patt. > "їм"
-ем	-ти	max. 3	2nd patt. > "ти"
-єм	-ти	max. 3	2nd patt. > "ти"
-им	-ти	max. 3	2nd patt. > "ти"
-їм	-ти	max. 3	1st patt. > "їм", 2nd patt. > "ти"
-ем	-ть	max. 3	
-єм	-ть	max. 3	
-им	-ть	max. 3	
-їм	-ть	max. 3	1st patt. > "їм"

Table 3: Surzhyk-specific rules on identification of the first-person plural verb in present tense.

The aforementioned rules have to respect a specific distance between the combinations of patterns; the second element of the combination has to be situated at a maximum distance of 3 words from the first element. Additionally, similarly to the general rules, these specific rules also have requirements to be fulfilled by the output word. For example, the fourth rule in the Table 3 requires the second element of the output that ends in "-їм" to be a word with a number of characters higher than the number of characters present in the word "їм." This type of requirement is also present in rules where it was necessary to prevent the false positive outputs that are personal pronouns of the third-person plural in dative case "їм." A similar requirement is present in rules where the output contains a final particle "-ти", according to this additional requirement the output has to be a word with a number of characters higher than the number of characters present in the word "ти." This excludes the false positive output "ти" as a single word expressing the personal pronoun of the second-person singular.

Initial Pattern "под-" Characterising the Verb Formation Process

In order to define verbs with a prefix "под-" that may be considered as Surzhyk lexicon, the pattern "под" has to be an initial part of the word. Additionally, the output word has to present a number of characters superior of 3, meaning it has to be a part of a verb and not a single preposition "под" that is composed of 3 characters. We may also define the personal pronouns that precede the verb, though in our samples their presence was limited; in most cases, personal pronouns were omitted or were expressed by a noun. Therefore, we decided to have general rules on the identification of verbs containing the prefix "под-." Based on this general rule, we can develop further restrictions in the future once we would have more digital material to analyse.

Patt.	Output word requests
под	Patt. > "под"

Table 4: Surzhyk general rule on the identification of the verbs with a prefix "под-".

Experimental Setting and Analysis of Results

In this section, we present the results for every combination of rules. The rules were implemented in the R programming language as a search of regular expressions in strings in a tidyverse fashion.[9]

During the validation phase, each rule was firstly tested on the previously analysed Surzhyk texts. Then, we tested the effectiveness of the rules on a new set of Russian texts. Since Surzhyk is a hybrid language between Russian and Ukrainian, and its elements basically come from Russian

9 Source code is available on Github for reproducibility: https://github.com/gmdn/Surzhyk.

and were imported in Ukrainian language, some of these elements may also be present in standard Russian.

For this reason, we included transcriptions of Russian audio samples to evaluate the effectiveness of the rules in terms of false positives (i.e. Russian texts classified as Surzhyk). We selected a subset of transcriptions of interviews present on the Radio svoboda (website Radio Liberty) ([29],[33],[34]).

We used the "true positive" or "false positive" measures to evaluate the effectiveness of the classification process. A "true positive" term is a term that corresponds to the aimed output criteria. For the identification of the verbs in the first-person plural in the present tense ending "-м," this means that the output entity has to be a Surzhyk verb in the first-person plural in present tense. When we refer to the identification of the verbs containing a prefix "под-," the output has to be a Surzhyk verb with a prefix "под-" and not, for example, a noun. Any non-Surzhyk term that is classified under these rules is considered a "false positive". The full documentation of the outputs and the R source code are available online for reproducibility purposes.[10]

Results of Surzhyk General and Specific Rules on Identification of the First-Person Plural Verb in Present Tense Applied on Surzhyk Corpus

The process of creating and testing the general rules may be considered a necessary part of this research that led us to the creation of specific rules. We firstly analysed the outputs for the combination of the final pattern "м" situated at a maximum distance 3 from the final pattern "ть." Then we checked the combination of the final pattern "м" located at a maximum distance 3 from the final pattern "ти." The output we received during the testing of these general rules on Surzhyk texts were mostly false positive and not related to the question of Surzhyk. Considering the low utility of the general rules (if compared to the specific rules), in this paper we decided to present the results of the outputs very briefly.

1) **ми + -м:** 10 results, 7 true positive.
The results that could be considered true positive are all verbs of the first-person plural in present tense with a non-standard ending "-м." The false positive results present a final pattern "-м" but are not verbs.

2) **самі + -м:** 2 results, 1 true positive.
The aimed output corresponds to a verb, while the false positive output is a noun.

3) **-м + -ти:** 5 results, 5 false positive.

10 https://github.com/gmdn/Surzhyk

All the outputs are not related to the question of Surzhyk; they pertain to Ukrainian standard language and are not verbs of the first-person plural. For these reasons, we decided to not present them.

4) **-м + -ть:** 39 results, 3 true positive.

The false positive results are 36. An analysis of the false positive results led to the conclusion that they contain the ending pattern "-ть," but do not actually correspond to the aimed search question of being verbs of the first-person plural. In the false positive outputs, some words emerged that were defined previously as false positive results; also, some verbs appeared with the standard ending of the third-person plural. Some of the outputs are standard Ukrainian elements, while others are new examples of Surzhyk. The emergence of the non-standard verb "єсть," for example, is repeated 5 times in different files. This verb is present also in the Russian language, written "есть" [jest'] and corresponds to Ukrainian "є" [je]. In this case, we define "єсть" [jest'] as part of the spoken Surzhyk corpora, due to the reasons connected to the record transcribing methods that were explained in the introduction and in the preparatory phase of the project. Other false positive results were considered as not important for the current research.

The specific rules regarding the identification of the first-person plural verb in Present tense lead us to the identification of 11 combinations of verbs with the first-person plural ending. Among a total amount of 12 outputs, only one result was a false positive. Since the specific rules are more detailed, they allowed us to have more precise outputs. We decided to present the outputs of the specific rules regarding the identification of the first-person plural verb in present tense in Table 5 referring to the rule.

Rule	File	1st word	2nd word	Text
ми + -ем	44	ми	кажем	Я з сім'ї двєнацить чоловік дітей, я восьма на щоту. Після мене ще брат і сестра, і менший самий брат, виродок ми на нього кажем.
ми + -єм	27	ми	працюєм	Сутки ми тут працюєм.
	27	ми	взнаєм	А ми взнаєм, я вже всіх знаю хто тут живе, ми ж вивчаємо жильців усіх по фамілії.
	31	ми	балакаєм	Та по п'ять, нечисто, ми чисто не балакаєм, хіба ми чисто балакаєм?
	31	ми	балакаєм	Та по п'ять, нечисто, ми чисто не балакаєм, хіба ми чисто балакаєм?
ми + -им	39	ми	бачим	Не знаю, може оце будуть вибори може шось воно буде лучше, хто його знає, уже і так пятнацить лєт пройшло, а ми лучшего на бачим, все хуже і хуже йде.

ми + -їм	39	ми	устроїм	Йдеш по городу в такий день, в отпуску, спрашують чо ти не на роботі, я говорю ето «в отпуску». Єслі не робиш, говоріт, то ми тебе устроїм.
самі + -ем	-	-	-	-
самі + -єм	31	самі	сієм	Самі сієм самі в'яжемо. Труд дуже важкий, коло його дуже багато роботи, пилява і все, але ше треба якось жити.
самі + -им	-	-	-	-
самі + -їм	-	-	-	-
-ем + -ти	-	-	-	-
-єм + -ти	-	-	-	-
-им + -ти[11]	-	-	-	-
-їм + -ти[12]	-	-	-	-
-ем + -ть	27	будем	злазить	Їдемо в автобусі, я вибачаюся, а я кажу «Де будем злазить?» Вона така в мене дуже мужня. Ето виходіть. А вона каже до мене «Мамо, не кричи».
	39	будем	возвращять	А сюди приїхав збиріженія у нас пропали. Не знаю… може слухаєш радіо, кажуть будем возвращять, а коли вернуть ніхто не знає.

11 Previously, when we did not specify that the second element has to follow the first one at maximum distance 3, and that the second output word containing the pattern "-ти" has to be longer than the pattern "-ти" in terms of characters, we obtained different results and they were false positives. The first combination of results was a Ukrainian demonstrative pronoun in instrumental case "цим," situated at position 93 and preceded by Ukrainian noun "мати" at position 92 in the file 44. The second combination of results was a Ukrainian noun, a toponym, "Рим" located at position 276 and preceded by Ukrainian personal pronoun "ти" at position 273 in the file 129. Both first elements of the two combinations were listed as false positive results after the first phase of project development, but with the improvement of the rules we managed to avoid these false positive outputs.

12 Notice that if checked singularly, the final pattern "-їм" is present once as part of the first-person plural Surzhyk verb "устроїм" (in file 39). Moreover, it may be identified three times as a Ukrainian personal pronoun in Dative case "їм" if we do not define that the output word has to contain more characters than the search pattern "-їм." The Surzhyk verb "устроїм" was already identified in the first phase of project development when searching for words ending in "-їм," and during the second phase while checking the rule that contained personal pronoun "ми" situated at distance 3 from verbs ending in "-їм."

-єм + -ть[13]	44	вообщєм	двадцять	Мені було двадцять три роки, як я вийшла заміж, ну вообщєм уже, да, двадцять три роки.
-им + -ть	39	мусим	ходить	Я даже не знаю, спрашивал одного, він каже: «та шоб дома не сидіть, то» каже, «мусим ходить на роботу.»
-їм + -ть	-	-	-	-

Table 5: The outputs list of Surzhyk specific rules on Identification of the first-person plural verbs in present tense applied on Surzhyk corpus.

Results of Surzhyk General Rule on Identification of the Verb with the Prefix "под-" Applied on Surzhyk Corpus

Testing the rule on the identification of the verbs containing a prefix "под-" gave us 8 outputs, of which 4 might be considered true positive; this means the output elements are verbs and под- is a prefix. As assumed, most of the false positive outputs are verbs with a prefix "по-" or, in the single case, a noun. The true positive outputs are verbs: "подожди," "подимаюся," "подработать" and "подвів."

Results of Surzhyk General and Specific Rules on Identification of the First-Person Plural Verb in Present Tense Applied on Russian Corpus

Surzhyk is a hybrid language between Russian and Ukrainian, which means that its element may also come from Russian. One of the patterns that is characteristic as a common final ending for Ukrainian, Russian, and Surzhyk words is the final pattern "-ем." Regarding the first Surzhyk element identification (verbs of the first-person plural in present tense), we firstly test the general rules and then the specific rules. Remember that for the first-person plural verbs in the present tense of Surzhyk, the specific rules were more effective compared to the general rules, as they reduced the number of false positive results.

During the testing process of the general rules on Russian corpus, we obtained 25 results; 20 of them were actually the outputs for the input search "-м" + "-ть" and 5 the results for the input "-м" + "ти." We did not analyse all the outputs in detail as it was concluded previously that the general rules were not effective for the identification of Surzhyk verbs of the first-person plural.

Secondly, it was decided to check if the specific rules may provide similar outputs. What was discovered by testing the specific rules was that the total number of matches was 10. As supposed, the highest number of matches (8 of 10) corresponded to the input combination containing the

13 The output is composed by two strings as required, but the first element is a Surzhyk adverb while the second one is a Ukrainian numeral; thus, they do not respect the aimed output and are to be considered false positive.

pattern "-ем." As we know, "-ем" is one of the possible common endings that are present in Russian and Ukrainian languages, and also in Surzhyk.

Analysing the results obtained with the specific rules on identification of the first-person plural verb in present tense, the outputs are as follows. Three of the specific rules on identification of Surzhyk verbs led to the identification of the elements in the standard Russian texts. Some of these elements were verbs with the infinitive ending "-ть," others were irrelevant results because of their affiliation to the categories of Russian nouns, adjectives, or personal pronouns that in different cases, such as dative, instrumental and locative, present the final pattern "-ем" or "-им."

Results of Surzhyk General Rule on Identification of the Verbs with the Prefix "под-" Applied on Russian Corpus

When we tested the rule on the identification of the verbs with the prefix "под-" on Russian corpus, we obtained 28 outputs. Among these results are present verbs with a prefix "под-," verbs with a prefix "по-," nouns with a prefix "под-," or nouns that have "под" as an initial part of the stem.

Discussion

Based on the results of the testing process, we can proceed with the discussion on the efficiency of our rules. Firstly, we discuss the results of the study on the identification of Surzhyk verbs of the first-person plural in present tense. The specific rules allowed to identify 11 combinations of verbs with the first-person plural ending. In a total amount of 12 outputs, only one result was false positive. 11 results obtained with the general rules corresponded to the results we had with the specific rules. The difference consisted in the number of false positive outputs that were limited to 1 by adoption of the specific rules. It can be concluded that the rules we provided were effective if applied to our texts; consequently, we list them for practical reasons and present the number of outputs for every combination.[14]

1st patt.	2nd patt.	Distance	Output word requests	Quantity of results	True pos.	False pos.
ми	-ем	max. 3		1	1	0
ми	-єм	max. 3		4	4	0

14 Do remember that all of these combinations have to respect the basic rule: the second element of the combination has to follow the first element at a distance from 1 to 3. In addition, the output words regarding the patterns "-їм" or "-ти" have to present more letters than the input pattern. This additional condition is to be applied on all the rules that present "-їм" or "-ти" in the first or second pattern of the combination.

ми	-им	max. 3		1	1	0
ми	-їм	max. 3	2nd patt. > “їм”	1	1	0
самі	-ем	max. 3		0	0	0
самі	-єм	max. 3		1	1	0
самі	-им	max. 3		0	0	0
самі	-їм	max. 3	2nd patt. > “їм”	0	0	0
-ем	-ти	max. 3	2nd patt. > “ти”	0	0	0
-єм	-ти	max. 3	2nd patt. > “ти”	0	0	0
-им	-ти	max. 3	2nd patt. > “ти”	0	0	0
-їм	-ти	max. 3	1st patt. > “їм” 2nd patt. > “ти”	0	0	0
-ем	-ть	max. 3		2	2	0
-єм	-ть	max. 3		1	0	1
-им	-ть	max. 3		1	1	0
-їм	-ть	max. 3	1st patt. > “їм”	0	0	0

Table 6: Summarising table of results. Verbs of the first-person plural in present tense

For true positive, we intended results that respected the aimed output of being verbs with the endings of the first-person plural in present tense or copula verbs in the first-person plural combined in the infinitive form. Though general rules can be seen only as a generalisation of the specific rules, and their outputs were mostly irrelevant in relation to our input and for the purpose of identifying Surzhyk verbs of the first-person plural, they allowed for the opportunity to find more Surzhyk words and thus means they may be used to enlarge the corpus of Surzhyk.

In regard to the second group of elements, the verbs containing prefix “под-,” we can see that the rule has to be improved for a better performance. However, testing this rule on Surzhyk corpus gave us several interesting verbs that do not appertain to Ukrainian nor to Russian vocabulary. Among the results were present verbs with a prefix “по-,” followed then by a letter “д” as an initial consonant of a stem or of another prefix, or nouns.

Patt.	Output word requests	Quantity of results	True pos.	False pos.
под	Patt. > “под”	8	3	5

Table 7: Summarising table of results. Verbs with a prefix “под-”.

In the final phase of project development, it was decided to verify whether the rules on identification behave as expected on new texts. We selected Russian texts of spoken language; in particular, some transcriptions of the real conversation and interviews. On one hand, one may suppose that no element should be identified as the rules we created are aimed at identifying Surzhyk elements. On the other hand, one may consider the possibility of having patterns we identified as Surzhyk within Russian text, especially the final pattern "-ем" or a prefix "под-," for example. We can assume that once we have a POST corpus of Surzhyk, we can develop the rules in accordance with the part-of-speech characteristics. These would allow for the reduction of inappropriate results, or even minimise the effectiveness of the Surzhyk identification rules when applied on Russian corpus. But we have to consider that Surzhyk is a mixing of Russian and Ukrainian standards, and for this reason, presenting elements of both languages is an important characteristic of its nature. This fact can explain why some rules on its identification are also efficient when applied on standard Russian texts.

Conclusion

The purpose of this research was to investigate whether it is possible to automatically identify the elements of a hybrid Ukrainian-Russian language Surzhyk. We focused our study on the particular group of Surzhyk-characterising elements we discovered in the analysed Surzkyk samples: the first-person plural verbs of Surzhyk in present tense and the Surzhyk verbs with a prefix "под-." The research was conducted with an example-based method. Through the creation of a written corpus, we studied the Surzhyk samples and defined the rules for pattern identification of hybrid verbs. The rules were then implemented and tested with the help of R.

By adopting an example-based method, we created rules on the automatic identification of the first-person plural verbs in present tense, and of verbs with a prefix "под-." We prepared a list of false positive results. For the automatic identification of verbs with a prefix "под-" we have one general rule, while for the first-person plural verbs identification we created two groups of rules: general and specific. The adoption of the specific rules led to the successful identification of the desired elements and reduced the number of false positive results we had with general rules.

The identified elements concerning verbs with the ending of the first-person plural were 11, while the ones regarding the verbs with a "под-" prefix were 4. The total amount of non-standard terms in the analysed texts were more than 1000.

At this point we are unable to provide an annotated corpus to the non-standard language in question, but we provide seven *Simplified terminological tables of Surzhyk* with a total amount of more than 1000 terms, a model of the deeper description of the hybrid language terms S*tandard terminological table of Surzhyk* with 5 entities, 16 specific rules on the automatic identification of the final pattern of the first person plural ending in present tense, and one general rule on the identification of Surzhyk verbs with a "под-" prefix implemented in R. Moreover, it was concluded that in order to improve the rules on Surzhyk automatic recognition, further studies

should develop towards Part of Speech Tagging (POST) applied on Surzhyk corpora. Finally, the study of the endings of the first-person plural of Surzhyk has demonstrated that there is an internal coherence in the Surzhyk verb phrase. Since this internal linguistic consistency is limited to our corpus, we propose to verify the hypothesis of internal coherence in Surzhyk verb phrase.

References

[1] Akademiia Nauk Ukrains'koi RSR. 1978. Istoriia ykriins'koi movy. Morfolohija. Kyiv: Naukova dumka.

[2] Akademija Nauk SSSR. 1960. Grammatika russkogo jazyka. Tom I. Fonetika i morfologija. Moskva: Izdatel'stvo Akademii nauk SSSR.

[3] Akademija Nauk SSSR. Institut russkogo jazyka. 1980. Russkaja grammatika. Tom I. Moskva: Izdatel'stvo Nauka.

[4] Allen, J. F. 2003. "Natural Language Processing." In Encyclopedia of Computer Science 4th, 1218-1222. UK: John Wiley and Sons Ltd. Chichester. dl.acm.org/citation.cfm?id=1074630.

[5] "An Introduction to R." In, The Comprehensive R Archive Network (CRAN). cran.r-project.org/doc/manuals/r-release/R-intro.html#The-R-environment.

[6] Appel, R., Muysken, P. 1987. Language Contact and Bilingualism. London: Arnold.

[7] Babych, N. D. 1993. Istoriia ukrains'koi literaturnoi movy. Praktychnyi kurs. L'viv: Vydavnytstvo Svit.

[8] Bakker, P. 1994. "Michif, the Cree-French mixed language of the Métis buffalo hunters in Canada." In Mixed languages: 15 Case Studies in Language Intertwining, Bakker P., Mous M. (eds.), 13-33. Amsterdam: Institute for Functional Research into Language and Language Use (IFOTT).

[9] Bakker, P. 1996. "Language intertwining and convergence: typological aspects of the genesis of mixed languages." In Sprachtypologie und Universalienforschung 49, 9-20.

[10] Bakker, P., Mous, M. 1994 "Introduction." In Mixed languages: 15 Case Studies in Language Intertwining, Bakker P., Mous M. (eds.), 1-11. Amsterdam: Institute for Functional Research into Language and Language Use (IFOTT).

[11] Barman, U., Das, A., Wagner, J., and Foster, J. 2014. "Code Mixing: A Challenge for Language Identification in the Language of Social Media. Proceedings of the First Workshop on Computational Approaches to Code Switching." Association for Computational Linguistics: 13-23. aclweb.org/anthology/W14-3902.

[12] Burlaka, T. K., Horpynych, V. O., Dudyk_ P. S., Taranenko, I. I., Pentyliuk, M. I., Tokar, V. P. 1993. Ukrains'ka mova. Chastyna 1. Za redaktsiieiu P. S. Dudyka. Kyiv: Vyshcha shkola.

[13] Buzuk, P. 1927. Istoriia ukrains'koi movy. Vstup. Fonetyka, Morfolohiia. Kyiv. Fotoperedruk z pisliaslovom Oleksy Horbacha.1985. München: Ukrains'kyi Vil'nyi Universytet.

[14] Chowdhury, G. G. 2003. "Natural Language Processing." Annual Review of Information Science and Technology 37: 51-89. onlinelibrary.wiley.com/doi/full/10.1002/aris.1440370103.

[15] Del Gaudio, S. 2010. On the Nature of Suržyk: A Double Perspective. München-Berlin-Wien: Wiener Slawistischer Almanach Sonderband 7.

[16] Feinerer, I., Hornik, K., and Meyer, D. 2008. "Text Mining Infrastructure in R." Journal of Statistical Software 25, 5. www.jstatsoft.org/article/view/v025i05.

[17] Filin, F. P. 1972. Proishozhdenie russkogo, ukrainskogo i belorusskogo jazykov. Istoriko-dialektologicheskij ocherk. M.-L.: Nauka.

[18] "Introduction to R Course." www.datacamp.com/courses/free-introduction-to-r.

[19] Kent, K. 2010. "Language Contact: Morphosyntactic Analysis of Surzhyk Spoken in Central Ukraine." LSO Working Papers in Linguistics 8, Proceedings of WIGL: 33-53. pdfs.semanticscholar.org/0705/54bea599cacdaac26bb3b753a0443a93026a.pdf.

[20] Kent, K. 2012. "Morphosyntactic analysis of Surzhyk, a Russian-Ukrainian mixed lect." PhD diss., University of Minnesota.

[21] Kuznecov, P. S. 1953. Istoricheskaja grammatika russkogo jazyka pod redakciej Avanesova R. I. Morfologija. Moskva: Izdatel'stvo Moskovskogo universiteta.

[22] Masenko, L. T. 2008. "Surzhyk: istoriia formuvannia, suchasnyi stan, perspektyvy funktsionuvannia." In Belarusian trasjanka and ukrainian surzhyk: Structural and social aspects of their description and categorization, edited by Hentschel G. and Zaprudski S., 1-37. Oldenburg: BIS-Verlag der Carl von Ossietzky Universität Oldenburg.

[23] Myers-Scotton, C. 2002. Contact linguistics: Bilingual encounters and grammatical outcomes. New York: Oxford University Press.

[24] Myers-Scotton, C. 2003. "What lies beneath: Split (mixed languages) as contact phenomena." In The mixed language debate, Matras, Y., Bakker, P. (eds.), 73-106. Berlin: Mouton de Gruyter.

[25] Olszański, T. A. 2012. The language issue in Ukraine. An attempt at a new perspective. Warsaw: Ośrodek Studiów Wschodnich im. Marka Karpia Centre for Eastern Studies.

[26] R Development Core Team "The R Base Package." stat.ethz.ch/R-manual/R-devel/library/base/html.

[27] R Development Core Team. 2013. "R: A Language and Environment for Statistical Computing." Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/.

[28] Roecker, S., Yoast, K., Wills, S., and D'Avello, T. 2018. "Introduction to R and RStudio." Statistics for Soil Survey. ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html#3_rstudio:_an_integrated_development_environment_(ide)_for_r.

[29] Rykovceva, E. 2018. "Za vremja pravlenija Putina samym effektivnym byla vojna." Radio svoboda, August 8, 2018. www.svoboda.org/a/29420499.html.

[30] Sanchez, G. 2013. Handling and Processing Strings in R. Berkley: Trowchez Editions.

[31] Silge, J., Robinson, D., 2008. Text Mining with R. A Tidy Approach. Online version: O'REILLY. www.tidytextmining.com.

[32] Ukrains'kyi pravopys. 2015. Kyiv: Naukova dumka, 108-117. www.litopys.org.ua/pravopys/rozdil2.htm#par80.

[33] Volchek, D. 2017. "Sirota pod statuej Stalina." Radio svoboda, May 6, 2017. www.svoboda.org/a/28470012.html.

[34] Volchek, D. 2018. "Smertel'nyj pomet baklanov. Pticy i ljudi v "Kislotnom lesu." Radio svoboda, August 20, 2018. www.svoboda.org/a/29437182.html.

[35] Wachsmuth, H. 2015. Text Analysis Pipelines - Towards Ad-hoc Large-Scale Text Mining. (Vol. 9383) Springer.