Ukrainian Redaction of Church Slavonic (URCS): Needs for Digitalization and Text Corpora Platform Generation. Part I.

Nataliya Popovych [1], Andriy Lutskiv[2], Oleksandr Mitsa [1], Olha Lyntvar[3] and Andriana Ivanova [1]

[1] *Uzhhorod National University, Narodna Square, 3, Uzhhorod, 88000, Ukraine*
[2] *Ternopil Ivan Puluj National Technical University, Ruska 56, Ternopil, 46001, Ukraine*
[3] *National Aviation University, Liubomyra Huzara Ave. 1, Kyiv, 03058, Ukraine*

Abstract

The article explores the issues of digitalization and possibilities of text corpus generation for the Ukrainian Redaction of Church Slavonic (URCS). This endangered language is still in use in Liturgical Services in certain regions of Ukraine (mostly Zakarpattia and Lviv Regions), as well as on bordering territories of Slovakia, Romania, and Poland. The given research is on its initial stage. It provides a brief overview of the URCS language history, and examines the interconnections between the Ukrainian language and URCS through examples from its usage in the modern Ukrainian language, as well as in Ukrainian literature of different genres and time periods. In addition, (4) the article suggests preservation and conservation approaches and strategies to URCS digitalization, including the creation of the text corpora platform, from both the user and developer perspectives, with the aim of ensuring its survival for future generations. The given article outlines a set of issues aimed at being solved through (1) the analysis and the classification of URCS text collections, (2) review of Ukrainian corpora and corpus tools available for Ukrainian speaking target users in the open access payment free as well as on reasonable fixed-price basis, (3) corpus-based analysis provided on the examples of URCS lemmas used in the modern Ukrainian language, in the texts of Ukrainian literature of different time periods as well as comparative analysis of URCS texts and their Ukrainian translations focusing on the accuracy, adequacy and faithfulness of specialized terminology and concepts. Furthermore, the article explores the potential of creating digital text corpora for URCS by utilizing modern technologies and methods by the interdisciplinary research team of linguists and IT experts. The creation of such a corpora platform could facilitate linguistic research, including studies on vocabulary, grammar, and syntax, as well as studies on the specific terminology, historical and cultural aspects of the language. The research also highlights the need for further collaboration among linguists, and digital experts to enhance the preservation and promotion of the URCS.

Keywords

Ukrainian Redaction of Church Slavonic (URCS), URCS text analysis platform (software), parallel corpus, comparable corpus, NLP, SketchEngine, mova.info, LancsBox 6.0.
URCS, URCS text analysis platform (software).

[1]COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: nataliya.popovych@uzhnu.edu.ua (N. Popovych); l.andriy@gmail.com (A. Lutskiv); alex.mitsa@gmail.com(A.Mitsa); olha.lyntvar@npp.nau.edu.ua (O. Lyntvar); andriana.ivanova@uzhnu.edu.ua (A.Ivanova);
ORCID: 0000-0001-6949-0771 (N. Popovych); 0000-0002-9250-4075 (A. Lutskiv); 0000-0002-6958-0870 (O. Mitsa); 0000-0003-4671-5514 (O. Lyntvar); 0000-0002-1733-4416 (A. Ivanova)

1. Introduction

The need for the creation of a URCS Text Analysis Platform (Software) that will preferably include parallel and comparable text corpora is urgent and significant for several reasons. Firstly, the URCS language is rapidly disappearing from current Church (liturgical) use. Secondly, there is constant planned substitution of it by the Russian Redaction of Church Slavonic in Ukraine (e.g. in Zakarpats'ka oblast') [1]. Finally, the preservation of the unique heritage of URCS in text and chant, which is still present in Church use now, is crucial for future generations.

The subject matter requires scientific exploration, analysis, and discussion due to the absence of adequately proven research results in the English-speaking scientific community and the lack of broad insight, deep comprehension or even solid awareness of URCS language issues within the global research community today. Despite the assiduous efforts of prominent Ukrainian scholars, such as Pavlo Hrytsenko and in particular Vasyl' Nimchuk, who demonstrated unwavering dedication to addressing URCS research problems, persistent issues remained unresolved in this field. Their work has remained unknown due to possibly weak information dissemination and a lack of research result sharing and subsequently adequate research result impact outside the Ukrainian-speaking scientific community. Nevertheless, Nimchuk's PhD students and colleagues continue effectively carrying out research in this area [2].

The research focuses on the following main issues:

1. The threat of extinction facing the URCS language due to its forced substitution by the Russian Redaction of Church Slavonic by Moscow authorities [3, 1].
2. The lack of evidence and research on the presence of URCS in Ukrainian literature of different time periods within English scientific literature, reviews, and articles on the subject matter.
3. The lack of evidence and research on the presence of URCS in modern Ukrainian language within English scientific literature, reviews, and articles on the subject matter.
4. The importance of digitalization to meet the needs and purposes of preserving URCS.
5. The main requirements of a URCS corpus or corpus-based tool.

Suggestions for the best approaches to digitalizing the URCS language, taking into account a vast majority of already existing corpora, are being developed to solve different specific tasks.

The outlined issues, which are aimed at being resolved, will require a series of analyses and reviews to be conducted in future research. These include: (1) analyzing and classifying URCS text collections, (2) reviewing Ukrainian corpora that are available for Ukrainian-speaking users in open access or on a reasonably fixed-price basis, (3) conducting corpus-based analyses on URCS lemmas in transliterated versions of URCS texts, modern Ukrainian texts of different genres and styles, and Ukrainian literature from different time periods. In order to meet the specific needs of researchers, it is necessary to develop a URCS query platform that includes parallel and comparable corpora of text and chant, which enable a comparative analysis of URCS texts and their Ukrainian translations. The development of such a platform should take into account the accuracy, adequacy, and faithfulness of the translation of URCS specialized terminology and concepts.

Digitalization and text corpus generation are essential for preserving the Church Slavonic of Ukrainian Redaction, which is an important aspect of Ukrainian cultural heritage and religious traditions. URCS is still used in Zakarpattia, Bukovyna and Lviv regions of Ukraine, where it has been preserved as a proof of its integral part of religious and cultural life in Ukraine for centuries.

In conclusion, creating a digital corpus of URCS would enable scholars and researchers to analyze its grammar (syntax), and vocabulary systematically. This would also facilitate the study of Ukrainian translations of URCS texts and allow for important amendments and corrections to be made.

2. Brief Historical Overview of the Church Slavonic of the Ukrainian Redaction (URCS): the Past and the Present

The URCS language has been the subject of extensive research by Academician V. Nimchuk and scientists at the Institute of the Ukrainian Language of the National Academy of Sciences of Ukraine (Institute of the Ukrainian Language NASU), resulting in numerous works in the Ukrainian language that explore its complexities and challenges.

The historical overview presented here is primarily based on the works of V. Nimchuk, N. Puriaeva, other scientists from the Institute of the Ukrainian Language at the National Academy of Sciences of Ukraine, H. Kuzems'ka, M. Skab, as well as M. Moser and the authors' own research. The authors have given due consideration and respect to other scholarly works in the Ukrainian language that focus on the history, functionality, development, and significance of the URCS language.

Many ethnic languages, such as Greek, Latin, Arabic, and Old Slavonic – Church Slavonic, have been elevated to the status of regional or world sacred languages. These sacred languages often cease to be used in daily communication and instead become reserved for cult purposes, acquiring the label of "dead languages" [3, 4].

The language traditionally known as Old Church Slavonic (OCS) was the language of religious practice in the territory of East Slavia from the 11th to 13th centuries. Prior to becoming the language of Kyivan Rus', the OCS language underwent three to four intermediate periods of its development [1, 2, 3, 4].

The first Old Church Slavonic texts were translated by St. Cyril and St. Methodius from Ancient Greek based on the language spoken by the Slavonic population of their native town Thessaloniki. When the saints arrived in Great Moravia, they had to adapt those texts to the local language spoken in the kingdom. As a result, the oldest monument of Old Church Slavonic, the Hlaholitic Leaflets, is characterized by the distinct use of the Czech and Slovak languages of that time. Then the brothers moved to Pannonia, where they had to adapt the Old Church Slavonic language for the Slavic population in that region as well. From there, the disciples of St. Cyril and St. Methodius spread the use of Old Church Slavonic to other regions in two directions: Croatia and the Bulgarian state [3,4].

Another significant period of Old Church Slavonic development occurred in the Bulgarian Empire, where it flourished and became enriched. Macedonia, which was part of the Bulgarian state and had important centers of book culture at that time, contributed to the language's enrichment too. Hence, the language obtained the colorful and distinct features of the Bulgarian and Macedonian language mix. This language, known as Old Bulgarian, was inherited by the first Eastern Slavic Church communities in the 10th century. After Christianity was introduced as the official religion in Kyivan Rus' in 988, Old Church Slavonic of the Old Bulgarian Redaction was spread throughout Rus'. However, it immediately began to be influenced by the communicative features of the Eastern Slavic language – the language of the local population. By the end of the 11th century, this language had acquired the common characteristics of the Eastern Slavic Redaction and was used as a language of the Church alongside the Old Rus' standard language [3, 4, 5, 6].

The East Slavic Redaction of Old Church Slavonic had distinct features of phonetics and morphology, and as books were edited in the capital of Rus', many colloquial words began to appear in the Old Church Slavonic texts, including liturgical ones. In particular, Old Kyiv words were frequently used in them.

Old Slavic texts were read in various regions of Kyivan Rus' according to the native pronunciation of the reader. The pronunciation of the capital city, Kyiv, which was a church center and metropolitan city, was regarded as the standard and as an exemplary. For instance, it is widely accepted that in the southern and southwestern regions of Rus', the letter "ѕ" was pronounced as a guttural sound similar to modern Ukrainian.

By the mid-13th century, the Old Slavic language had transformed into a variety that was typical of the Ukrainian language in the Kyivan state. This version of the East Slavic variant of the Old Slavic language was in use from the mid-12th century to the end of the 13th century, and since that time and later its Redactions are referred to as Church Slavonic by philologists. While modern-day Russia and Belarus have gradually developed their own Redactions of the Church Slavonic language, orthoepy of the capital and metropolitan Kyiv exerted significant influence and authority in these regions. Evidence

of this can be seen in the liturgical orthography of Russian Old Believers in the north of Russia today, who still use the letter "ѣ" as a back-palatal fricative, which partially corresponds to Ukrainian, and also pronounce hard consonants before "e," similar to the Ukrainian pronunciation [3, 5].

Starting from the end of the 18th century, two redactions of the Church Slavonic language coexisted in various Ukrainian denominations: the Old Kyiv or the Church Slavonic language of Ukrainian Redaction in the Greek Catholic Church, which was displaced from Right-Bank Ukraine to the territory of the Austrian Empire and continued to develop as the Ukrainian Greek Catholic Church (UGCC) in 1795, and the Old Moscow Redaction in the Orthodox Church on the territory of Ukraine. This determined the further development of the Ukrainian liturgical language in these denominations. In the UGCC, the liturgical Church Slavonic language was never used as an instrument of national assimilation of Ukrainians. The preservation of the Ukrainian pronunciation allowed it to be perceived as a chronological (old Ukrainian), functional (church, as opposed to secular) and stylistic (highly literary, as opposed to everyday spoken language) variant of the Ukrainian language. The URCS language was perceived by Greek Catholics as the language of their native faith, rite, and therefore, their native (not foreign) language [5, 6, 7, 8, 9].

2.1. The URCS CS Language Today

Following the forced displacement of the autochthonous Ukrainian Redaction of the Church Slavonic language by the Russian Redaction, there are now only few remaining regions in modern Ukraine where this language is still actively used in liturgical practice. These regions likely include three oblasts of Ukraine, namely Lviv, Zakarpattia, and Chernivtsi.

Within the Lviv oblast, the Univ Lavra of Holy Ascension (UGCC) is considered to be the primary center for the continuous and constant use of the Ukrainian Redaction of the Church Slavonic language [10].

The Ukrainian Greek-Catholic Church (UGCC) on the whole has a natural inclination to preserve the Ukrainian Redaction of the Church Slavonic language in its liturgical practices, given that it is the official liturgical language of this Church. This language, being the liturgical matrix of the UGCC, forms the basis for all other liturgical translations, even though they are now used more frequently than the URCS language [11].

The URCS language is used as a main liturgical language alongside modern Ukrainian in the Greek Catholic Diocese of Mukachevo in Zakarpatska oblast [12].

2.2. Main Features of URCS Pronunciation

Thus, we may claim that every CS redaction has their own orthoepy, which significantly impacts the way the text is transliterated [7]. By transliteration we mean the transfer of a text lettering into the target alphabet. Hence, it looks more like a reproduction of a language at a phonetic level preserving its characteristic features.

Every Slavic Church is also reflected differently in terms of phonetic representation. Regardless of the place where theologians-writers and composers created their sacred-language texts – Bulgaria, Romania, Slovenia, Ukraine or Russia, every people is bound to transliterate their texts following the language norms of their land [4, 5, 13]. Ukrainian pronunciation is engraved in the world famous “Grammar” by archbishop Meletius Smotryts’kyi, which fixed some foreign and later language forms though, still preserved Ukrainian stress and traditional pronunciation of letters: *г, ѳ, ѣ, и.* argued that *ѣ* was pronounced as *і* and not *є* (according to Russian tradition) [14].

This specific Ukrainian phonetics was preserved regardless of the numerous tsarist and synodal decrees up to the end of the eighteenth century when following the Valuev Circular (1863) and Ems Ukaz (1876) the ruinous attack on Ukrainian orthoepy was launched. The language was humiliated, nicknamed ‘distorted Russian’, ‘twisted Polish’, ‘language of the lowest societal layers’[13]. But all this did not prevent the Ukrainian people from praying in the language of their ancestors. Since our aim is to highlight the importance of preserving old Ukrainian tradition of liturgical texts, we, following V. Nimchuk, claim that Ukrainian redaction of Church Slavonic, which was mainly revealed through

Ukrainian orthoepy at a desolate time of lacking Ukrainian statehood acted as one of those spiritual and cultural factors that guarded the integrity of Ukrainians as an ethnic group. This redaction contributed to the formation of a single cultural and linguistic area, was a characteristic feature of the Ukrainian Church, Ukrainian identity [3, 4, 13].

To conclude, we want to stress the importance of preserving the Ukrainian redaction of Church Slavonic as a source of Ukrainian national identity, whose linguistic and cultural heritage was appropriated by Russian church culture which allowed them to distort both Ukrainian realities, create myths and build their own deceitful linguistic, historical and cultural background.

3. Ukrainian Corpora and Corpus Platforms

In a previous publication, we presented a classification and overview of corpora. In this research, we use a three-fold classification of corpora that includes content-based corpora and corpus tools, functional annotation set and aim-based corpora, and generation-based corpora [14].

- Content-based corpora and corpus tools
- Functional annotation set and aim-based corpora
- Generation-based corpora [14]

Content-based corpora and corpus tools can be further categorized into national, professional, parallel, comparable, specialized, and task-based (adaptable or mixed) [14]. Among the various corpus platforms available to Ukrainian users, we focus on two projects: the Corpus of the Ukrainian Language developed by N. Dartchuk, O. Siruk, M. Langenbach, Ya. Khodakivska, and V. Sorokin at the Institute of Philology of TKU in Kyiv [15] and the Laboratory of Ukrainian and the General Regionally Annotated Corpus of Ukrainian (GRAC) [16]. These projects are among the most developed of the Ukrainian corpora and corpus tools [14].

Mova.info is a corpus platform of the Ukrainian language, which allows users to search and analyze a large collection of Ukrainian texts. The platform contains a diverse range of texts from different genres and time periods, including literary works, scientific papers, news articles, and more. Users can search for specific words or phrases, view concordances and collocations, and perform various types of linguistic analysis.

We conducted a small experiment to explore the use of URCS words in Ukrainian literature. The experiment was carried out using the mova.info corpus platform. Out of 100 randomly selected URCS words, 80 were found to be used in literary works within the corpus. This experiment highlights the connection between URCS and the literary language of Ukraine.

GRAC is the Corpus of the Ukrainian language, which counts 1.875 billion tokens in its 16 version.

SketchEngine is another multilingual text analysis software, which provides corpora in 14 languages, including Ukrainian. The Ukrainian corpus is presented as ukTenTen – Ukrainian corpus from the web [17].

The Ukrainian Web Corpus (ukTenTen) is a Ukrainian corpus made up of texts collected from the Internet. The corpus belongs to the TenTen corpus family which is a set of web corpora built using the same method with a target size 10+ billion words. Sketch Engine currently provides access to TenTen corpora in more than 40 languages. Data for the Ukrainian Web 2020 corpus consists of texts from May 2014 and July–August 2020. The Wikipedia part is from December 2020. The final size of the corpus contains 2.5+ billion words [17]. There are 3,282,586,754 tokens, 2,592,516,436 words, 129,751,817 sentences, as well as 7,204,875 web pages. The Ukrainian Web 2020 corpus is lemmatized by CSTLemma and part-of-speech tagged by RFTagger using two different tagsets (MULTEXT-East Ukrainian PoS tagset, which is more-detailed and Universal Dependencies PoS tagset showing only basic parts of speech) [17].

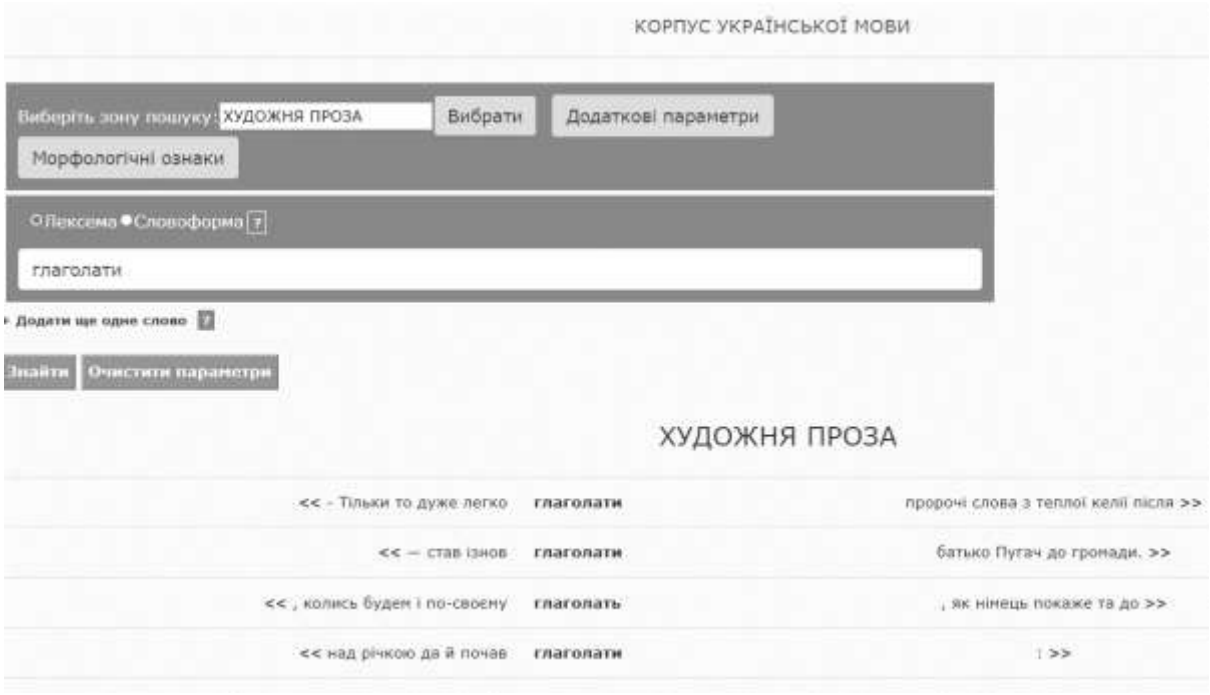

Figure 1: Concordance Search Query Result for URCS Lemma “hlaholaty” in Ukrainian Literature Corpus on mova.info.

A complete set of Sketch Engine tools is available to work with this Ukrainian Web corpus to generate

- keywords– terminology extraction of one-word units
- word lists – lists of Ukrainian words organized by frequency
- n-grams– frequency list of multi-word units
- concordance – examples in context
- text type analysis – statistics of metadata in the corpus [17].

4. Developing the URCS Corpus Platform

The development of the URCS corpus platform must take into account the previous approaches as well as the needs and expectations of users. Therefore, it is important to consider the user side during the development process.

Firstly, the URCS platform must be user-friendly, take into account different types of users (ordinary people interested in the URCS vocabulary and texts, medium level experts, and linguists) and be easy to navigate.

Secondly, the platform should offer a diverse collection of URCS texts, encompassing liturgical, literary, and historical works, which can be subjected to lemmatization and morphological analysis. Users should have access to a broad range of texts for linguistic research purposes. These texts are typically printed in the URCS alphabet. However, one issue that arises is the lack of consistency in the use of this alphabet in such publications.

Thirdly, the platform should allow for easy downloading and exporting of texts in various formats, such as OCR-ed PDF or TXT and others as it shown on Figure 1. This will allow users to conduct their own analysis and research of uploaded corpus or corpora and download the received results.

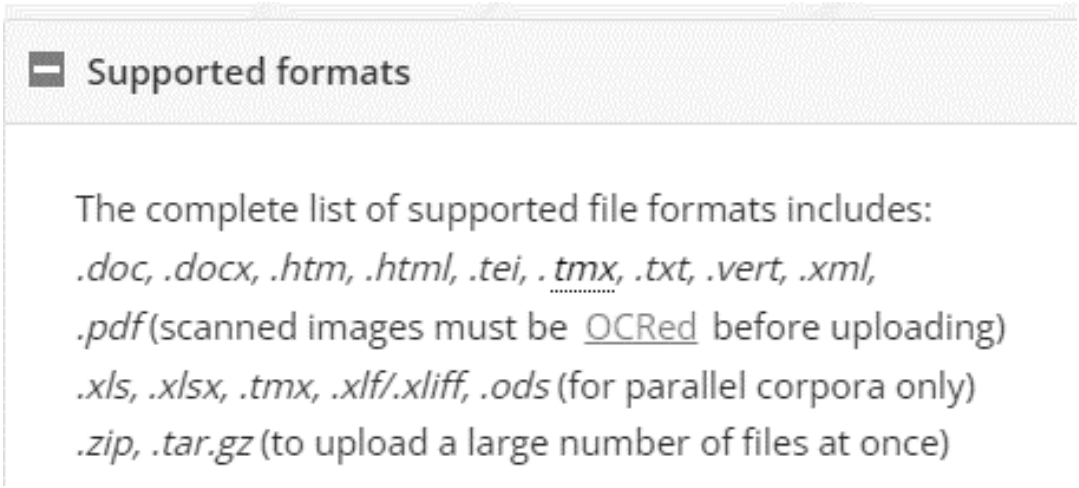

Figure 2: Options and Supported Formats on SketchEngine Platform

Unfortunately, all scanned, but not OCR-ed pdf texts are not supported by the platform. In this case neither the user, nor the corpus will benefit. Users will not be able to submit their own texts or suggest corrections to existing texts.

Finally, the platform should provide support and resources for users who may not be familiar with the URCS language or digital corpus research. This can include user guides, tutorials, and a help desk for technical support as it is provided by many corpus tools like LancsBox 6.0 and its former versions and corpus platforms like SketchEngine or ГРАК-16 [16, 17].

4.1. Previous Backgrounds in Development

Section 4.1 presents examples of team work results (on the material of Bible books in different languages), which serve to illustrate the developmental context and potential solutions that the team may employ for generating the URCS text corpora platform. In our previous team publications, the solutions for *adaptable text corpus development for specific linguistic research* were suggested [14]. It was also described the *effectiveness of automated linguistic analysis using a big data approach* [18, 19, 20]. It is worth providing here its *data processing workflow implementation* [14] and computational experiments [18].

According to the nature of input data we used approaches for Big Data processing, so software should fulfill these requirements. Developed corpus tool prototype was based on software components of Apache Hadoop ecosystem (Hortonworks Data Platform 3.1). Suggested corpus tool was implemented basing on Lambda architecture.

Application was developed with Java 8 programming language and Spring Framework 5. Workflow was implemented with Apache Spark 2.3 [21] components: stages of workflow implemented as a Stanford Core NLP Pipelines in Apache Spark SQL using Spark Datasets which were well supported in Java. Pipelines were implemented by using appropriate libraries.

Apache Tika with TesseractOCR used for Data ingestion of source data in binary formats (images, raster and vector PDFs, DOC, DOCX). Bliki-core and edu.umd.cloud9 libraries used for handling Wikipedia's tags.

Main workflow steps 1-9 provided by Stanford Core NLP [22] library and LangTool. Nowadays LangTool has the best support of Ukrainian language for purposes of POS-tagging and text spellchecking. Workflow step 10 was implemented with Apache Spark MLLib: TF-IDF calculation, matrix operations in LSA. Workflow step 11 was implemented with Word2VecfJava library [23]. Steps 13 and 14 were implemented with available API of Wikipedia, Dictionary of Ukrainian Language [24], Oxford Dictionaries API [25] and Glosbe API [26]. Building inverted index was done with Apache Lucene Library. Index was stored in Apache Solr which was embedded into HDP 3.1 platform. Vocabulary of gathered metadata from step 12 stored into RDBMS PostgreSQL 9.6 to minimize time of data access.

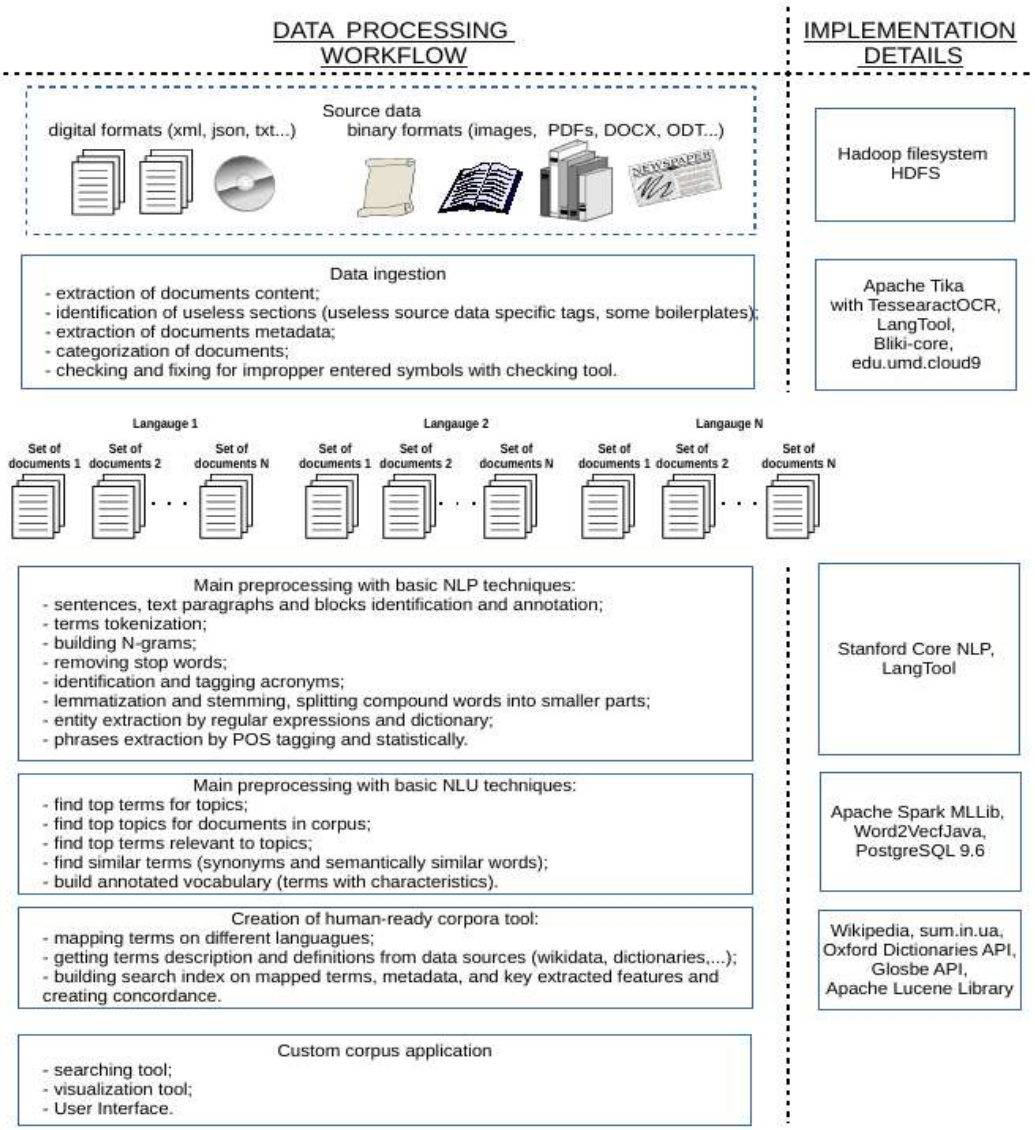

Figure 3: Data Processing Workflow

Computational experiments carried out on Wikipedia dumps and open text documents of Ukrainian and English texts. In the implementation process free or/and open source software were used. Data source were open or free of charge [14].

For computational experiment corpora of different editions, translations and languages of the Bible were compiled to verify the suggested approach. For experiment there were taken the books which were used for mobile applications in SQLLite format and were imported into RDBMS PostgreSQL to work with Apache Spark [18, 19].

Due to that investigation every Bible edition was treated as a subcorpus, i.e., a set of chapters. Each chapter had its own sentences and terms. After ETL the most important keywords, term POS-tags, relations between terms and other features for each chapter were obtained. After statistical processing each subcorpus (book) and its documents (book chapters) had its own characteristics. Some books were logically subdivided into stories, but the number of stories depended on translation and varies from zero to 1252 and other books contained less number of translated books. Due to those two factors and in order to provide more accurate results it was suggested to divide each book into documents by chapter criterion and to prepare custom ETLs for different types of books. After ETL there were to be done the following processing steps: sentence and word tokenization, calculation of terms frequencies, finding collocations (N-grams with high probabilities), POS-tagging, stop words filtering, lemmatization,

calculation of TF-IDFs, building of term-document matrix with TF-IDFs, SVD with obtaining low-dimensional term-document matrix representation. Developed adaptable corpus also allowed to choose the custom k-value [18, 19].

Two other successful experiments were conducted, one focused on the effectiveness of automated linguistic analysis using a big data-based approach, while the other developed an adaptable corpus translation module [18, 19, 20, 27].

The developing team has also an extensive experience in developing software to meet the needs of cultural sector. In particular, the professionally developed map of Ukrainian dialects [28] displays the settlements where the entered word is still being used. The map contains a big amount of dialect data (more than 32 thousand words) and enhances language diversity preservation. The client part of an interactive map is created with the help of the library React.js, programming language JavaScript and the library of managing the sate MOBX. The server part of the informational system is written in the programming language Ruby using framework Ruby on Nails. Relational DBMS Postgresql has been used as the primary database along with Redis cache for caching some of the most frequently used data. One more development [29] is connected to displaying the toponyms taken from web-portal “Diia” developed by Ministry of Digital Transformation of Ukraine and with help of data provided by the Institute of National Remembrance. The processed data have been visualized with the help of cartographic web-service Google Maps. Every decommunized object got a pin on the map. The interactive map has been integrated to the web portal Analitycs-UA.

4.2. User Interaction Processes and Technologies’ Overview

In the end of this section we provide a brief overview of the user interaction processes and technologies used as well as technical details related to the implementation.

The project contains an authentication flow that is designed to allow users to securely log in to our platform and access their accounts. After successful authorization, the user has access to the main feature which is the creation of the corpora, their viewing and interaction with it. The system provides two options for corpus creation: manual text formation or based on text file generation. Furthermore, there are options for editing texts on the selected pages in the corpus and viewing experience that allows users to read content as a physical book.

The project uses several technologies to deliver a seamless experience. The system uses Next.js for the front end and Nest.js for the API side. Next.js is a React-based framework that allows us to build powerful and performant user interfaces for the platform. On the back end side, Nest.js and PostgreSQL are used to manage server-side platform operations. Also being a cloud provider AWS S3 feature, it is highly scalable and secure object storage, designed to store and retrieve any amount of data from API and it is used for storing blob text files. The project leverages cutting-edge technologies and architectures to deliver a top-notch user experience.

4.3. Search Query Experiments (SketchEngine)

A good example of user-friendly software is SketchEngine, which we tested by creating our own URCS corpus. Following a clear instruction [30], one of the URCS text in PDF was uploaded to the platform (scanned images have been OCRed before uploading). For our search query experiments we use the prayer book "Promin dushi" edited by Greek Catholic Diocese of Mukachevo [31]. The text is written in the URCS language, but transliterated using the letters of the modern Ukrainian alphabet based on the phonetic principle ("write as you hear"). The only difference between this URCS text and modern Ukrainian writing is the presence of stress mark on almost every word, except for function words (stop words). The example of one search query result can be seen on Figure 4.

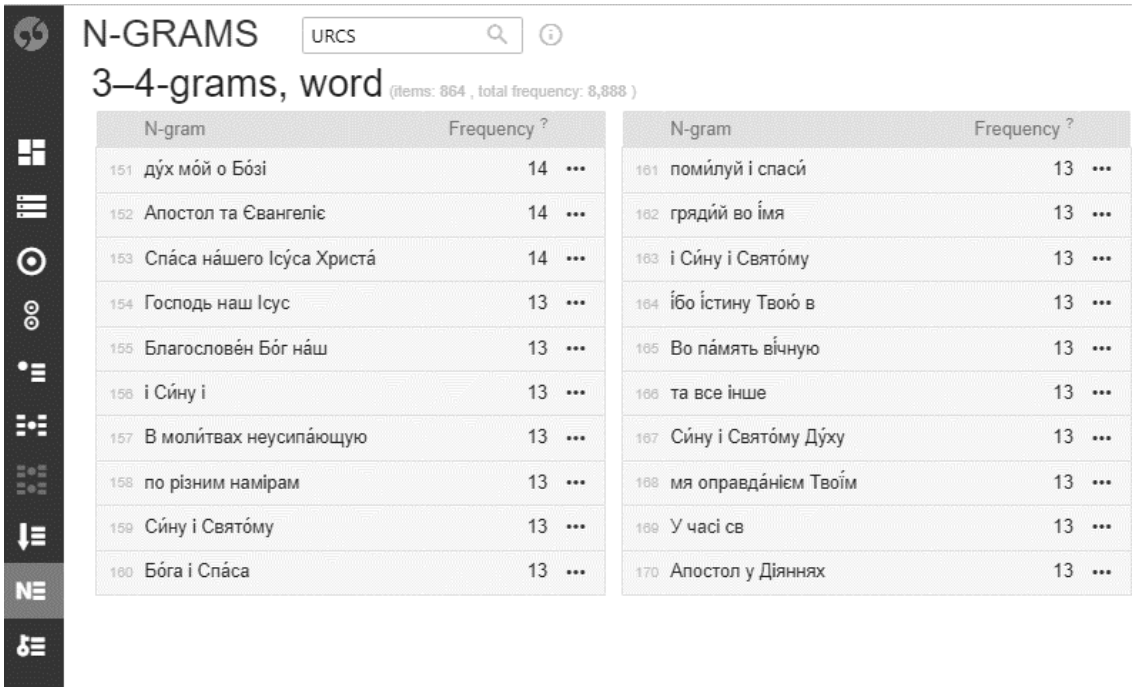

Figure 4: 3-4-Ngrams, URCS Corpus, Uploaded to SketchEngline Platform

Given the diverse origins of URCS texts in terms of time and place of publication, a standardized alphabet and consistent use of diacritical marks poses a significant challenge for the development of the URCS corpus platform. The only solution might be to segment the texts into separate corpora based on their historical periods. This approach would enable scholars and researchers to analyze and compare different versions of the URCS language through creating parallel as well as comparable corpora which will also help tracing the development of its linguistic features over time.

ПСАЛОМЪ 103.

Благословѝ дшѣ моѧ ГДа,* ГДи Бѣ҃е мо́й, возвели́чилсѧ є҆сѝ зѣлѡ.
Во и҆спов҃ѣда́нїе и҆ въ велелѣпотѣ҃ о҆блѣклсѧ є҆сѝ,* сдѣ́йсѧ свѣ́томъ ꙗ҆кѡ ри́зою.
Простира́й нѣ́ко ꙗ҆кѡ ко́жѹ,* покрыва́й вода́ми превы́спренняѧ твоѧ.
Полага́й о҆́блаки на восхожде́нїе твоѥ́,* ходя́й на крилу́ ве́трею.
Творя́й а҆́нгелы твоѧ дѹ́хи,* и҆ слу́ги твоѧ пла́мень о҆́гненный.
Основа́й зе́млю на твёрди є҆ѧ́,* не преклони́тсѧ во ве́къ ве́ка.
Бе́здна ꙗ҆кѡ ри́за о҆дѣ́нїе є҆ѧ́, на гора́хъ ста́нѹтъ во́ды.
Ѿ запрещенїѧ твоегѡ побѣгнѹтъ, ѿ гла́са гро́ма твоегѡ ѹ҆бою́тсѧ.

Псалом 103

Благослови́ душе́ моя́ Го́спода: / Го́споди Бо́же мо́й, возвели́чился е́сй зіло́.
Во ісповіда́ніе і в велелі́поту облёклся е́сй, / одія́йся сві́том я́ко ри́зою.
Простира́яй не́бо я́ко ко́жу, / покрива́яй вода́ми превы́спренняя своя́.
Полага́яй о́блаки на восхожде́ніе свое́, / ходя́й на крилу́ ві́треню.
Творя́й а́нгели своя́ ду́хи / і слу́ги своя́ пла́мень о́гненний.
Основа́яй зе́млю на твёрди е́я, / не преклони́тся в ві́к ві́ка.
Бе́здна, я́ко ри́за одія́ніе е́я, / на гора́х ста́нут во́ди.
От запреще́нія Твоего́ побі́гнут, / от гла́са гро́ма Твоего́ убоя́тся.

Figure 5: Psalm 103 in URCS, and in URCS using the Stress Marked Ukrainian Transliteration

Despite the successful creation of the URCS corpus on the SketchEngine platform, the search results obtained are insufficient. Only the ngram search query produces accurate results while all other search functions fail to work or display inaccurate outcomes. The reason behind this issue is the mismatch

between the grammar and stylistics of our URCS corpus and the Ukrainian language, resulting in only transliterated text that uses the modern Ukrainian alphabet being matched by the search functions.

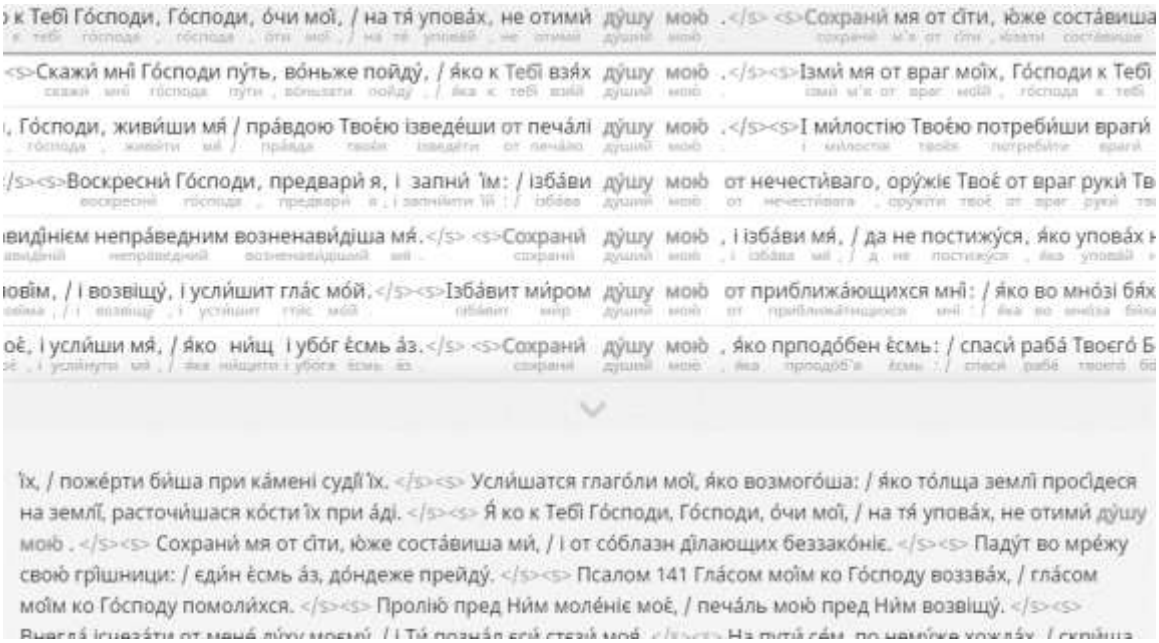

Figure 6: POS-tagging of URCS Search Query Showing the Wrong Initial Grammatical Forms

As we can understand, the search functions are likely not able to accurately analyze the linguistic properties of the text.

Solution 1: Preprocessing the URCS text may facilitate the recognition of the linguistic features of the original language, leading to more accurate search results. However, the feasibility of this solution is contingent upon the availability of tools and resources for giving equal value to the transliterated text grammatically and stylistically as original URCS text, as well as the compatibility of the transliterated URCS text with SketchEngine's search functions.

Solution 2: An alternative approach is to utilize a different platform or tool that is better suited to the URCS language and corpus. The selection of an appropriate tool may require careful evaluation of its capabilities in handling the specific linguistic properties of the Church Slavonic language of the Ukrainian Redaction.

Solution 3: To ensure the preservation and accessibility of the extant URCS texts, it is imperative to create curated collections of the texts and develop a tailored corpus platform that can accommodate the unique linguistic features of the URCS language and the specialized search queries required for meaningful research. Such a custom platform would require significant resources and expertise, but it would enable scholars and researchers to conduct more accurate and targeted analyses in the URCS corpora, thus contributing to a deeper understanding of this historically significant language.

By taking into account the user side during the development of the URCS corpus platform, the developer team should ensure the needs and expectations of users and provide (1) a valuable resource for linguistic research and (2) preservation efforts.

The main expectations of the user are as following.

- different types of text data ingestion (URCS of different publishing periods)
- text processing
- semantic tagging of each part of a corpus (e.g. UCREL Semantic Analysis System [32])
- qualitative and quantitative analyses which are based on different statistical characteristics
- comparison with different translations of the same text and map terms in different languages
- texts which will be stored and analyzed in the corpus should be chosen only by linguists to build proper dependencies and lead to proper statistics to prevent side effects in statistics
- linguists can choose proper calculation methods of text preprocessing and analysis and these methods should be customizable [19, 20]

5. Conclusions

The article highlights the significance of preserving the Ukrainian Redaction of Church Slavonic language (URCS) through digitalization and the creation of text corpora. URCS is an endangered language that is still used in liturgical services in certain regions of Ukraine and neighboring countries. The article provides a brief history of URCS and explores its connection with the Ukrainian language. It also reviews Ukrainian corpora and corpus tools, and analyzes URCS lemmas using corpus-based techniques. To preserve the URCS language, the article suggests various preservation and conservation approaches, such as creating a URCS corpora platform or a separate corpus tool like LancsBox.

The platform would facilitate linguistic research on vocabulary, grammar (syntax), specific terminology, and historical and cultural aspects of the language. Additionally, the development of a URCS query platform that includes parallel and comparable corpora of texts is necessary to meet the specific needs of researchers.

The article identifies several issues that need to be addressed in the future, including the morphological analysis and classification of the URCS text collections, opportunities for lemmatization and text elaboration, as well as creating various search queries for users.

Overall, the article stresses the significance of preserving the Ukrainian Redaction of Church Slavonic language (URCS) as a cultural and linguistic heritage of Ukraine for future generations.

URCS is not only an important liturgical language alongside modern Ukrainian, but also a valuable source of linguistic, historical, and cultural knowledge. Therefore, the article proposes digitalization and corpus-based research as a means to conserve and promote the language, and to ensure its transmission to future generations.

6. References

[1] V. Nimchuk, Yakoyu movoyu molylasya davnya Ukraina, Video, 2012. URL: https://www.youtube.com/watch?v=V_Lhfv5J4WA&t=334s&ab_channel=brownianbox

[2] “Izbornyk. Istoriia Ukrainy IX-XVIII st. Pershodzherela ta interpretatsii”, (2003). URL: litopys.org.ua

[3] V. Nimchuk, Ukrains’ka mova – svyashchenna mova, Liudyna i svit (1992), 11–12, 28–32.

[4] V. Nimchuk, Literaturni movy Kyivs’koyi Rusi, Istoriya Ukrains’koyi kul’tury, 1 (2003). URL: http://litopys.org.ua/index.html

[5] V. Nimchuk, Leksyka davn'orus'koi movy, Istoriia ukrains'koi movy: Leksyka i frazeolohiia, (1983), 29—163.

[6] V. Nimchuk Davn'orus'ka spadschyna v leksytsi ukrains'koi movy, Kyiv, (1992).

[7] M. Moser. Cerkovnoslov’jans’ka mova ukrajins’koji redakciji v dzerkali mizhnarodnoji slavistyki. Balcania et Slavia, 2, 2022, 133-142.

[8] M. Skab, Mova Tserkvy v Ukraini kintsia XX – pochatku XXI st. yak chynnyk formuvannia natsional'noi svidomosti, Bohoslovs'kyj visnyk (2013), 8, 8-16.

[9] N. Puriaeva, Ukrayns'ka mova v liturhijnij praktytsi ukrayns'kykh tserkov, Problemy humanitarnykh Nauk, Seriya «Filolohiya», (2018), 42, 128-146.

[10] Univ Lavra of Holy Ascension (UGCC), (2007). URL: http://studyty.org.ua/index.php?option=com_files&Itemid=52;

[11] Shevchuk S., Church Slavonic is the Official Liturgical Language of the UGCC (in Ukrainian), 2020. URL: https://synod.ugcc.ua/data/glava-ugkts-tserkovnoslovyanska-mova-ofitsiynoyu-liturgiynoyu-movoyu-ugkts-4276/

[12] The Greek Catholic Eparchy of Mukachevo. (in Ukrainian), updated 2023. URL: https://mgce.uz.ua/

[13] H. Kuzems'ka, Yakoiu movoiu molylasia davnia Ukraina: Pravyla ukrains'koi transliteratsii tserkovnoslov'ians'kykh tekstiv, Kyiv, KZhD ”Sofiia”, 2012.
[14] Meletii Smotrytskyi. Hramatyka / Pidhotovka faksymilnoho vydannia ta doslidzhennia pamiatky V. V. Nimchuka. — K.: Naukova dumka, (1979), 111. 492 . (Faksymile).
[15] N. Dartchuk, O. Siruk, M. Langenbach, Ya. Khodakivska, and V. Sorokin, Corpus of the Ukrainian Language (Ukrainian) (2023). URL: http://www.mova.info/corpus.aspx?l1=209
[16] General Regionally Annotated Corpus of Ukrainian (GRAC) (Ukrainian, English), (2023). URL: http://uacorpus.org/Kyiv/ua
[17] SketchEnengine platform, (2023). URL: https://www.sketchengine.eu/
[18] A. Lutskiv, N. Popovych, Big data-based approach to automated linguistic analysis effectiveness, Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP, Lviv, (2020), 438-443.
[19] A. Lutskiv, N. Popovych, Big data approach to developing adaptable corpus tools CEUR Workshop Proceedings, Lviv, (2020) 374-395.
[20] A. Lutskiv, N. Popovych, Adaptable Text Corpus Development for Specific Linguistic Research, Proceedings of IEEE International Scientific and Practical Conference Problems of Infocommunications. Science and Technology, Kyiv, (2019), 217-223.
[21] O. Levy, Dependency-Based Word Embeddings, 2014, URL: https://www.aclweb.org/anthology/P14-2050
[22] Stanford CoreNLP 3.9.2 (updated 2018-11-29). URL: https://corenlp.run/
[23] R. M. Reese, A. S. Bhatia, Natural Language Processing with Java, 2 nd ed., Birmingham: Packt Publishing, (2018), 318.
[24] Academic Dictionary of the Ukrainian Language, 2018. URL: http://sum.in.ua/
[25] Oxford Dictionaries API, 2023. URL: https://developer.oxforddictionaries.com/
[26] Glosbe API, 2023. URL: https://glosbe.com/a-api
[27] A. Lutskiv, R. Lutsyshyn, Corpus-Based Translation Automation of Adaptable Corpus Translation Module, CEUR Workshop Proceedings, Lviv, (2021), 2870, 511–527.
[28] O.V. Mitsa, H.V. Shumytska, V.V. Sharkan, N.F. Venzhynovych &H.I. Dulishkovych, Interactive map of dialects as the professional training tool for philology students, Information Technologies and Learning Tools, vol. 88, no. 2, , (2022), 126–138. doi: https://doi.org/10.33407/itlt.v88i2.4787
[29] M. Lupei, M. Shlahta, O.Mitsa, Y. Horoshko, H. Tsybko & V. Gorbachuk, Development of an Interactive Map Within the Implementation of Actual State and Public Directions, in 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), IEEE, (2022), 384-387.
[30] Create a new corpus from files, 2023. URL: https://www.sketchengine.eu/guide/create-corpus-from-files/#toggle-id-1
[31] A. Solans’kyi, Promin dushi, Uzhhorod, (2017), 830.
[32] The UCREL semantic analysis system, (2023). URL: https://www.researchgate.net/publication/228881331_The_UCREL_semantic_analysis_system
[33] V. Brezina, P.Weill-Tessier, & A.McEnery, #LancsBox v. 5.x. [software], 2020. URL: http://corpora.lancs.ac.uk/lancsbox.
[34] L.Bilen’ka-Svystovych, N. Rybak, Tserkovnoslovianska mova. Pidruchnyk zi slovnykom, 2012.
[35] H. P. Klimchuk, Cerkovnoslov’jans’ki zapozychennja v publicystyci Mikhaila Grushevs’kogo, Filolohichni studii, 2009, 3, 53-64.