DIACU: A dataset for the DIAchronic analysis of Church Slavonic

Maria Cassese[α], **Giovanni Puccetti**[α], **Marianna Napolitano**[β], **Andrea Esuli**[α]
[α] Istitute of Science e Tecnologies of Information "A. Faedo" (ISTI-CNR)
{name.surname}@isti.cnr.it
[β] University of Modena and Reggio Emilia, Foundation for Religious Sciences
napolitano@fscire.it

Abstract

The Church Slavonic language has evolved over time without being formalized into a precise grammar. Therefore, there is currently no clearly outlined history of this language tracing its evolution. However, in recent years, there has been a greater effort to digitize these resources, partly motivated by increased sensitivity with respect to the need to preserve multilingual knowledge. To exploit them, we propose DIACU (DIAchronic Analysis of Church Slavonic), a comprehensive collection of several existing corpora in Church Slavonic. In this work, we thoroughly describe the collection of this novel dataset and test its effectiveness as a training set for attributing Slavonic texts to specific periods. The dataset and the code of the experiments are available at https://github.com/MariaCassese/DIACU.

1 Introduction

The diachronic development of Church Slavonic has not been comprehensively codified in a unified historical grammar, which makes it particularly interesting for linguistic research. There are relevant studies that allow us to trace its evolution through textual variants, regional redactions, and shifts in orthographic and lexical conventions over time (Eckhoff and Janda, 2014; Tomelleri, 2022; Ferro et al., 2018).

However, identifying the regional and chronological influences on the linguistic features of specific texts is still a challenging task. This issue becomes particularly significant in the case of doctrinal and liturgical texts in Church Slavonic, which, initially translated from Greek into a language specifically regulated and constructed for ecclesiastical knowledge, was gradually transformed under the influence of Slavic culture and languages.

To study these phenomena, we created a large-scale dataset of texts in the Church Slavonic language, accompanied by chronological and geographic annotations. In total, we collected 652 documents from 4 different language variants.

This collection can serve two separate purposes. First, it can serve as a unified corpus for linguists and humanities scholars to investigate diachronic language phenomena manually. Second, it can be used as a training set for machine-learning-based attribution methodologies.

2 Related Works

Ancient languages can be analyzed in their spatial and temporal evolution. In particular, Old Church Slavonic is a language that has experienced a non-linear evolution over time. Born as an ecclesiastical language transplanted among a people and a region (the Great Moravia), it has undergone orthographic, lexical, and morphosyntactic variations — both unintentional during the copying process and deliberate ones, with editions rendered into Slavic vernacular languages.

When considering the spatial variation of a language, in Natural Language Processing, we typically refer to language identification methods, which are approaches aimed at identifying regional variants of the same language or languages that share a common proto-language. A representative work in this field is that of Wu et al. (2019), in which the authors train an SVM for all the language identification tasks of the VarDial Evaluation Campaign from 2016 to 2019. The tasks included the identification of similar languages and dialectal variants.

Regarding temporal evolution, NLP methods can support both the study of classical philology, by providing computational tools for the analysis of ancient texts (Bamman and Burns, 2020; Bamman and Crane, 2011), and the improvement of recognition systems for historical languages (Celano, 2020). Although limitations exist for all ancient languages, work carried out in Latin and Greek

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 101–107

	Century & Language				
	Old Church Slavonic 9th - 11th	Church Slavonic 12th - 17th	New Church Slavonic 18th	Ruthenian 15th - 18th	
Name	N. Docs	N. Docs	N. Docs	N. Docs	Total
Cyrillomethodiana	3	132	3	-	138
Syntacticus	6	33	-	-	39
Old Russian Hagiographic Literature	-	15	10	-	25
Russian Language National Corpus	-	106	-	-	106
Ruthenian Corpus	-	-	-	344	344
Total	9	286	13	344	652

Table 1: The composition of the DIACU dataset, subdivided into datasets and ages.

(Celano et al., 2016) is far more frequent. This is not the case for Old Church Slavonic, as there are still few studies on diachronic variation in Old Church Slavonic using NLP methods. One of these is the article by Lendvai et al. (Lendvai et al., 2025), in which a large collection of texts in Old Church Slavonic and Old East Slavic was digitized to evaluate the impact of the sentence segmentation on retrieval performance. Given a text available in both language variants, they developed a benchmark dataset aligned at the lexical and subsentential levels. The results showed that, for this task, classical similarity-based models still outperform large language models. For this work, two datasets were collected. The ground truth dataset consists of two versions of the Life of Paul and Juliana: one in Old Church Slavonic, extracted from the Codex Suprasliensis (10th century), and one in Old East Slavic, contained in the Great Menaion Reader (GMR, 16th century). In contrast, the test dataset consists of the March volume of the Great Menaion Reader. Equally relevant is the work of 2023 by Lendvai et al. (Lendvai et al., 2023), in which a dataset including six diachronic and cross-linguistic variants of Slavic Pre-Modern language is created. The six datasets span the period from the 10th to the 18th century and include different genres and language variants. This dataset was created to investigate the capabilities of the BERT model in classifying historical religious texts as a domain adaptation task by fine-tuning on masked language modeling.

3 The DIACU Dataset

The need for the DIACU dataset arises from the challenges faced by historical languages, which suffer from a limited amount of available textual resources, mainly due to a) the lack of digitized collections and, when such collections do exist, their dispersion across various digital libraries and portals; b) the difficulty of accessing manuscript collections from specific and limited geographical areas; c) the challenge of defining a short chronological boundary as an additional criterion, alongside the geographical one. An additional challenge includes the digitalization of works using HTR or OCR technologies (Scherrer et al., 2018; Pedrazzinia, 2020; Lendvai et al., 2024). Recent studies in this field have shown that the limited availability of critical editions of historical Church Slavonic texts passed down through manuscript tradition significantly slows progress and requires substantial correction efforts for model training. In the last instance, a unified literary standard is lacking, even in the case of available digital editions, resulting in a high percentage of orthographic and linguistic inconsistencies within the corpora.

The DIACU (DIAchronic analysis of Church Slavonic) dataset includes five collections of texts (Cyrillomethodiana: uni-sofia.bg, Syntacticus: syntacticus.org, Old Russian Hagiographic Literature: spbu.ru, a part of the National Corpus of the Russian Language (RNC): ruscorpora.ru, and a sample of the Ruthenian Corpus: UD_Old_East_Slavic-Ruthenian).

Among the datasets considered for the construction of DIACU, the RNC emerges as the most relevant. Its significance is primarily due to the 2020 expansion, which introduced new annotated data through the development of the Rubic model (Lyashevskaya et al., 2023). This improved parsing and lemmatization, particularly on historical and non-standard texts (Savchuk et al., 2024). In our dataset, we included a subset of this corpus, comprising 106 Old East Slavic documents from the Middle Russian Corpus, and released as part of the Universal Dependencies Treebank starting from UD v2.4 [1].

[1] https://github.com/UniversalDependencies/UD_

Language	Bulgaria	Poland	Ukraine	Russia	Latvia	Serbia	Turkey	Greece	Italy	Egypt	Syria	Belarus	Unknown	Total
OCS	5	1	–	–	–	–	–	–	–	1	–	–	2	9
CS	48	–	2	144	1	26	4	3	1	–	1	–	56	286
NCS	1	–	–	–	–	–	–	–	–	–	–	–	12	13
Rut	–	–	15	–	–	–	–	–	–	–	–	–	329	344

Table 2: Distribution of languages across regions.

Language	Old Russian	East Slavic	Old Serbian	Middle Bulgarian	Resavski	New Bulgarian	Middle Russian	Old East Slavic - Belarus	Old East Slavic - Ukraine	Not specified	Total
OCS	–	–	–	–	–	–	–	–	–	9	9
CS	55	1	18	56	25	5	118	–	–	8	286
NCS	–	–	–	–	–	–	–	–	–	13	13
Rut	–	–	–	–	–	–	–	329	15	–	344

Table 3: Distribution of languages across historical and regional variants.

The corpus is subject to access restrictions; therefore, we were able to use only its publicly shared section.

Another part of the dataset consists of the Ruthenian Treebank, containing 344 texts written in *prosta mova* (*ruska mova,* Old Belarusian, Old Ukrainian). This sample of legal and non-fiction texts, dated approximately between 1380 and 1650, is drawn from the Ruthenian Corpus, a historical language resource currently under development by an independent research consortium. Within DIACU, we have included the texts covering the period from the 14th to the 18th centuries. This decision, although the Ruthenian language never became a liturgical language, stems from the project's overarching aim: to create a diachronic dataset that may serve as a reference for tracing linguistic and literary variation over time and across regions.

The Cyrillomethodiana web portal constitutes another resource, incorporating 138 texts of Bulgarian origin, spanning various genres from the 10th to the 18th century (Totomanova, 2021), which brings together several projects contributing to the Histdict system [2] and related digital tools.

DIACU also includes texts from Syntacticus, an umbrella project that brings together the PROIEL Treebank, the Tromsø Old Russian and Old Church Slavonic Treebank (TOROT), and the ISWOC Treebank (Information Structure and Word Order Change in Germanic and Romance Languages). These resources all share a unified annotation system and common linguistic priorities. From this resource, DIACU integrates 39 texts (Berdicevskis and Eckhoff, 2020).

In addition, texts from the Old Russian Hagiographic Literature dataset, available on GitHub [3], were included. The texts considered comprise the *List of Lives*, a collection of 25 hagiographic texts dating from the 15th to the 17th centuries. The library also provides a tool for lexical research through concordances, which can be used by installing Old Russian fonts.

Overall, the documents in DIACU cover the period from the IX to the XVIII century, corresponding to four linguistic variants: **Old Church Slavonic** (OCS): 9th – 11th century; **Church Slavonic** (CS): 12th – 17th century (with different revisions: Bulgarian, East Slavic, Serbian); **New Church Slavonic** (NCS): 18th century; **Ruthenian**, ruska mova (Rut): 15th – 18th century.

Table 1 shows the subdivision of DIACU into sources and periods: OCS, CS, NCS, and Rut. In total, there are 652 documents with varying numbers written in each linguistic variant: 9 in OCS, 286 in CS, 13 in NCS, and 344 in Rut. This linguistic classification follows the standard definitions adopted in both international and Italian Slavic studies (Garzaniti, 2019). The languages mentioned refer to the sacred written languages used in Orthodox Slavic countries, except for the Ruthenian, which was never officially recognized as a liturgical language, but functioned instead as a medium of religious communication directed toward the lay population (Nedeljković, 2011).

Each document includes its title, language, regional language variant, and the region of origin of the edition. Some titles were in Old Slavonic or Russian, while others were in English. To standardize the information across documents, we included both the original title and its scientific transliteration, and, when available, the Latin title as well. The data concerning the region of origin are presented in Table 2. It can be noted that most of the OCS documents come from Bulgaria, the region where the disciples of Cyril and Methodius resorted after being expelled from the Great Moravia. The

Old_East_Slavic-RNC

[2] https://www.resilience-ri.eu/news/in-our-service-catalogue-histdict/

[3] https://github.com/vintagentleman/scat-content

Setting	Base			DRO			
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Support**
New Church Slavonic	0.900	0.692	0.783	0.833	0.769	0.800	13
Church Slavonic	0.951	0.958	0.955	0.954	0.951	0.953	286
Old Church Slavonic	0.750	0.333	0.462	0.800	0.444	0.571	9
Ruthenian	0.971	0.988	0.980	0.968	0.983	0.975	344
Macro avg	0.893	0.743	0.811	0.889	0.787	0.834	652
Weighted avg	0.958	0.960	0.958	0.956	0.957	0.956	652

Table 4: Results of the classification in the Base and Distributional Random Oversampling (DRO) settings.

only exceptions are the Codex Suprasliensis, from Supraśl in Poland, and the Psalterium Sinaiticum, held in the Monastery of Saint Catherine in the Sinai Peninsula in Egypt. The class CS is the one with the most regional variety. Most of the documents come from Russia (144), followed by Bulgaria (48) and Serbia (26). Additionally, there are documents from other regions, including Greece, Turkey, Ukraine, Latvia, and others. Finally, the Ruthenian class includes documents from Belarus (329) and Ukraine (15). This distribution is confirmed by the second table 3, where the historical and regional language variants are shown. The OCS documents attested in the present collection are written in the Old Bulgarian recension of the language. In the case of the CS texts, in line with the regional variation, a prevalence of documents in the Russian (Old and Middle Russian, 55 and 118), Bulgarian (Middle Bulgarian, 56), and Serbian (Old Serbian and Resavski, 18 and 25) variants is observed, followed by a spurious minority of other varieties.

3.1 Challenges

Collecting texts from different digital sources poses many challenges. In addition to those discussed in Section 3, one major issue is the presence in the dataset of Private User Area (PUA) characters. Their presence does not hinder the classification of texts by historical periods, since they can easily be included in the feature extraction process. However, collecting the appropriate fonts is essential to ensure the correct visualization of the overall textual content extracted from various webpages, each requiring different character sets and fonts. Building on this, we are working on an interface for the correct visualization of the entire dataset and we will develop a character mapping between PUA and Unicode code points to unify the characters across the whole dataset.

4 Classification of Church Slavonic Variants

As a case study for the use of the DIACU dataset, we train a machine-learning-based classifier that attributes Slavonic texts to different periods of time. We train a predictor in a 4-class classification setting, including all the ages available in DIACU. As the machine learning algorithm, we used a logistic regressor, as it often proved to be one of the most effective and efficient algorithms for text classification (Pranckevičius and Marcinkevičius, 2017), and it is beyond the scope of this first work to investigate fine-grained optimization of machine-learning methods for this dataset. Indeed, this classification task aims to validate the dataset rather than to provide a tool for the temporal attribution of Slavonic texts.

The logistic regressor is trained on basic stylometric features that do not overlap with semantic features: token length, number of characters per sentence, part-of-speech n-grams, character n-grams, and syntactic dependency n-grams, the latter three all with unigrams, bigrams, and trigrams.

Before extracting the stylometric features, the text is preprocessed. Since the documents are transcriptions of manuscripts, they contain substitutions of letters or missing words with a variable number of dots and sections of text enclosed in parentheses. To improve the text quality we remove the following patterns from the documents: (1) numbers smaller than 1000 (to be sure that no dates or other relevant numbers were involved); (2) numbers followed by letters indicating paragraphs (e.g. 242v); (3) biblical references or hymns; (4) dots in square brackets; (5) symbols for division into verses or paragraphs and other noisy symbols; (6) square brackets around one- to three-letter sequences. Square brackets surrounding larger portions of text were kept in place because they often abutted words that were not otherwise separated by

punctuation. Lastly, the text is lowercased.

The documents are divided into segments ranging from 8 to 400 tokens. This is done because the texts vary in length significantly. Both whole documents and segments are encoded as TF-IDF–weighted vectors. Training is carried out using a *leave-one-out* approach in which each document is tested using the set of remaining documents as training, which makes the results statistically more robust than other protocols, such as k-fold.

To mitigate the unbalanced number of documents in the four classes in DIACU, we compare two settings: a standard classification one, which we call BASE, and one with oversampling to balance the classes, which we call DRO from the name of the oversampling algorithm we use: Distributional Random Oversampling (DRO). DRO creates random synthetic samples of the minority class in the training set by leveraging the distributional patterns of words from the original documents (Moreo et al., 2016).

The DRO algorithm has two hyperparameters: the number of features to retain (it was settled on 80% after testing 80% and 100%); and the new proportion of the minority class examples versus the majority class in the synthetic data generation. We tested 20%, 50%, and 80%, and the best F1 was achieved by balancing the data to have an equal number of training examples (50% ratio) among the two classes[4] For details on the method, refer to (Leocata et al., 2025), the base for this work.

4.1 Results

Table 4 reports the scores achieved by our classifiers. In the Base setting, we see high Precision ($\geq$ 0.9) in the two most represented classes (CS and Ruthenian) and one of the less represented ones, NCS. On the contrary, precision in OCS is lower (0.75). Similarly, recall is higher for the two most represented classes. Among the two least represented ones, NCS also has a lower value, 0.69, and OCS is even lower, 0.33. As a result, F1 shows a similar pattern where the classifier achieves a per-class F1 higher than 0.95 on both CS and Rut, and lower scores for NCS 0.78 and OCS 0.46.

The DRO setting shows scores following a similar pattern as the BASE setting, where more populated classes are better identified than least populated ones, as expected. However, through DRO, the overall Macro Average F1 score rises from 0.81 to 0.83, and specifically, the per-class F1 score in NCS goes from 0.78 to 0.8 and OCS from 0.33 to 0.44, with negligible F1 losses in CS and Rut. The weighted average F1 score suffers a minor decrease, but we remark that for the leave-one-out setting, Macro Average is most appropriate.

5 Conclusions and Future Works

In this work, we collect a new dataset, DIACU, based on aggregating existing resources for Church Slavonic texts. The dataset is composed of 652 documents divided into 4 linguistic variants: Old Church Slavonic, Church Slavonic, New Church Slavonic, and Ruthenian.

As a first test case for the dataset, we evaluate its usability as a training set for machine-learning-based approaches to the attribution of Church Slavonic texts to different periods of time, and find that it enables the development of effective models, achieving F1-scores above 80%.

However, DIACU does not overcome some of the limitations inherent in processing historical-language texts that span such an extensive chronological range. One of the main limitations lies in the scarcity of texts belonging to the OCS and NCS categories, corresponding respectively to the first (OCS) and the most recent phase (NCS) of the considered periodization. Moreover, the dataset contains noisy elements such as diacritics, ligatures, graphic variants, and paragraph markers. We partially removed them through the pre-processing phase, but there is still room for improvement. A final relevant issue concerns the editions used: these have not been compared to the original manuscripts. As a result, potential editorial errors or inconsistencies in the criteria adopted by different editors are also reflected within DIACU.

Future expansion of the dataset will include a larger number of texts for each historical phase and a more detailed analysis of editorial criteria . Additional texts from other sources are expected to be included, originating from projects currently under development. In particular, the RNC Corpus of Birchbark Letters [5], made publicly available after May 2025. Another direction for future work is to incorporate a larger number of OCR and HTR-processed texts and to provide direct links to the digitized manuscripts.

[4] In the four classes case the classifier is built training four one-vs-the-rest binary classifiers, and assigning the class with the highest score. DRO is applied to each one-vs-rest classifier.

[5] https://github.com/UniversalDependencies/UD_Old_East_Slavic-Birchbark/tree/master

6 Acknowledgments

This work was supported by the PNRR (National Recovery and Resilience Plan) project Italian Strengthening of ESFRI RI Resilience (ITSERR) founded by the European Union—NextGenerationEU (CUP:B53C22001770006).

References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *Preprint*, arXiv:2009.10053.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, page 1–10, New York, NY, USA. Association for Computing Machinery.

Aleksandrs Berdicevskis and Hanne Eckhoff. 2020. A diachronic treebank of Russian spanning more than a thousand years. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5251–5256, Marseille, France. European Language Resources Association.

Giuseppe G. A. Celano. 2020. A gradient boosting-Seq2Seq system for Latin POS tagging and lemmatization. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–123, Marseille, France. European Language Resources Association (ELRA).

Giuseppe G. A. Celano, Gregory Crane, and Saeed Majidi. 2016. Part of speech tagging for ancient greek. *Open Linguistics*, 2(1).

Hanne M Eckhoff and Laura A Janda. 2014. Grammatical profiles and aspect in old church slavonic. *Transactions of the Philological Society*, 112(2):231–258.

Maria Chiara Ferro, Laura Salmon, and Giorgio Ziffer. 2018. *Contributi italiani al XVI Congresso Internazionale degli Slavisti:(Belgrado 20-27 agosto 2018)*. Firenze University Press.

Marcello Garzaniti. 2019. *Gli slavi: storia, culture e lingue dalle origini ai nostri giorni*, 2 edition, volume 207 of *Manuali universitari. Lingue e letterature straniere*. Carocci.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2023. Domain-adapting BERT for attributing manuscript, century and region in pre-Modern Slavic texts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 15–21, Singapore. Association for Computational Linguistics.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Retrieval of parallelizable texts across Church Slavic variants. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 105–114, Abu Dhabi, UAE. Association for Computational Linguistics.

Piroska Lendvai, Maarten van Gompel, Anna Jouravel, Elena Renje, Uwe Reichel, Achim Rabus, and Eckhart Arnold. 2024. A workflow for HTR-postprocessing, labeling and classifying diachronic and regional variation in pre-Modern Slavic texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2039–2048, Torino, Italia. ELRA and ICCL.

Martina Leocata, Alejandro Moreo, and Fabrizio Sebastiani. 2025. The *Questio de aqua et terra*: A computational authorship verification study. *Preprint*, arXiv:2501.05480.

Olga Lyashevskaya, Ilia Afanasev, Stefan Rebrikov, Yana Shishkina, Elena Suleymanova, Igor Trofimov, and Natalia Vlasova. 2023. Disambiguation in context in the russian national corpus: 20 years later. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*, pages 307–318, Moscow, Russia. Dialogue2023 Conference.

Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. 2016. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 805–808, New York, NY, USA. Association for Computing Machinery.

Olga Nedeljković. 2011. The linguistic „diglossia" of gavrilo stefanović venclović and „prosta mova" in the literature of the orthodox slavs. *Serbian Studies Research*, 2(2):7–80. Napomene i bibliografske reference uz tekst.

Nilo Pedrazzinia. 2020. Exploiting cross-dialectal gold syntax for low-resource historical languages: Towards a generic parser for pre-modern slavic. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Svetlana O. Savchuk, Timofey Arkhangelskiy, Anastasiya A. Bonch-Osmolovskaya, Ol'ga V. Donina, Yuliya N. Kuznetsova, Ol'ga N. Lyashevskaya, Boris V. Orekhov, and Mariya V. Podryadchikova. 2024. Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34.

Yves Scherrer, Achim Rabus, and Susanne Mocken. 2018. New developments in tagging pre-modern orthodox slavic texts. *Scripta & e-Scripta*, 18:9–33.

Vittorio S. Tomelleri. 2022. *When Church Slavonic meets Latin. Tradition vs. innovation*, pages 201–232. De Gruyter Mouton, Berlin, Boston.

Anna-Maria Totomanova. 2021. Electronic research infrastructure for bulgarian medieval written heritage: history and perspectives. *Diacronia*, (14):1–9.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. Association for Computational Linguistics.