Towards improving OCR accuracy with Bulgarian Language Resources

Ivan Kratchanov[1][0000−0002−0430−7953], Laska Laskova[2][0000−0002−6931−9082], and Kiril Simov[2][0000−0003−3555−0179]

[1] Digitization Centre, National Library "Ivan Vazov", Plovdiv, Bulgaria
ivankra@gmail.com

[2] AIaLT, Institute of Information and Communication Technologies, Sofia, Bulgaria
{laska|kivs}@bultreebank.org

Abstract In 2017, the National Library "Ivan Vazov"–Plovdiv, embarked on a digitalization project whose ultimate purpose is to provide both learners and scholars with several types of content, including periodicals and books published during the late Bulgarian National Revival and afterwards, in the decades before the communist era (1870s-1940s). We focus on the technical aspects of the digitalization project that involves optical character recognition (OCR) and requires proper handling of Cyrillic texts. The paper provides insight into the library's joint activities with its partners from the Institute of Information and Communication Technology at the Bulgarian Academy of Sciences to develop relevant tools and methodologies, by stressing the mutual benefits from the co-operations. The library's participation in the project CLaDA-BG, integrated within the European CLARIN and DARIAH infrastructures, offered a chance to take advantage of the multidisciplinary expertise of the partnering organisations and to develop the best methodology for OCR and consequently to enhance the methods of using and handling the acquired machine-readable text.

Keywords: Digitization · Cultural Heritage · Digital Library · Optical Character Recognition · Spelling Models · Modern Bulgarian

1 Introduction

The paper discusses the current efforts of the National Library "Ivan Vazov"–Plovdiv (NLIV) in making digitized content accessible to learners and scholars and focuses on the technical aspects of a digitalization project that involves optical character recognition (OCR) and requires the proper handling of Bulgarian Cyrillic texts, especially texts published before the last major spelling reform from 1945 (historical texts). It provides insight into the library's joint activities with the Institute of Information and Communication Technology at the Bulgarian Academy of Sciences (IICT-BAS) for the development of relevant tools and methodologies.

Our goal is twofold: (1) to perform a correct OCR on historical texts and (2) to normalize them, i.e. to convert various old spellings to the present one. The first step is essential for the publication of the original documents (newspapers, magazines, books, etc.). The latter is important at least for two reasons: it makes possible for users to search a corpus of both historical and preset-day documents with a query input in the current Bulgarian orthography and allows for the application of NLP tools built for contemporary Bulgarian on the normalized texts.

In order to achieve our first goal, we planned several experiments. The best option is, of course, to train a professional OCR software to perform OCR tasks for old Bulgarian orthography in the best possible way. Thus, our first experiment was to train the ABBYY FineReader system on a lexicon provided by the IICT-BAS group. The lexicon is a version of the contemporary inflectional lexicon of Bulgarian. It contains word forms converted in accordance with the writing rules of an old spelling. The conversion was based on rules that take into account the combination of letters in the word form, their position and some relevant grammatical features. As a result, the new "old" version of the lexicon contains 1 121 872 word forms. After the training of ABBYY FineReader, we performed evaluation on the basis of a scanned version of all issues of the "Science" magazine published in 1881, a total amount of 5485 running words. The percentage of non-recognized words dropped from 4.9% to 4.4%. The number of non-recognized hyphenated words per page was reduced from 6.9 to 5.55. Although these results are not significant, they show that training with knowledge resources is possible and that has the capacity to improve the result from OCR.

2 The Problem: Spelling Variations and Old Orthography Models in Bulgarian Printed Historical Texts

Optical recognition and access to texts printed before the last orthographic reform of the Bulgarian language (1945) is of utmost importance for any researcher in social sciences and humanities, whose work is related to 18^{th}–19^{th} century Bulgaria. The reform known as the Fatherland's Front Reform, has brought about the current rendition of the language written and spoken by Bulgarians today. Before 1945, there were several attempts at creating an exhaustive set of orthographic prescriptions (models) for written modern Bulgarian as opposed to the example of Church Slavonic.

Among those models, some proved to have more impact than others [3, 7]: the Drinov model (1870–1899), its slightly modified version, the Drinov-Ivanchev model (1899–1921), the short-lived Omarchevski model (1921–1923) and an updated version of the Drinov-Ivanchev orthography (1923–1945). They were developed by various authorities—writers, educationalists, scientific organizations, such as the Bulgarian Literary Society (BAS predecessor), or specially appointed committees—and for all of them, there were several topics of major importance:

- modification of the Old Bulgarian alphabet in order to have an adequate representation of the modern Bulgarian phonemes. For instance, the exclusion/inclusion of the letter **щ** from the alphabet was a subject of ardent discussion. While **щ** represents the consecutive pronunciation of the sounds /ʃ/ and /t/, each of them has its own letter, **ш** and **т**, respectively. Some argued that **щ** should be replaced by the combination of **ш** and **т**.

- mapping of sound changes onto letters. For example, in modern Bulgarian, the sound /ɨ/ represented by the letter **ы**, has reflected in /i/ that is already represented by the letter **и**, thus rendering **ы** redundant. From phonological point of view, keeping **ы** and several other redundant letters (**ѣ**, **ѧ**, **ѫ**, **ѩ**, **ѭ**, **і**, **ꙗ**) in use was meaningless, but in the times when Bulgarian identity was being (re)built, many considered those letters an evidence and a symbol of cultural continuity.
- selection of regional phonomorphological norms as the basis for the creation of a standard language. Different dialects offered different solutions. One question that remained open for decades because of the substantial linguistic variation related to origin, concerned the spelling of endings for 1st and 2nd conjugation present verbs in first-person singular and third-person plural, for example ***въpвя*** [vɤr'vʲɤ], '(I) am going' and ***въpвят*** [vɤr'vʲɤt], '(they) are going'. Depending on their region of origin and/or considerations about the prestige associated with some of the vernaculars, authors of various prescriptive texts suggested different spellings. If the inflectional inventory of the dialect included only the "hard endings" [ɤ]/[ɤt], the letter **а** seemed to be the most appropriate choice: ***въpва*** [vɤr'vɤ], ***въpват*** [vɤr'vɤt]. The "soft endings" [ʲɤ]/[ʲɤt] were represented in accordance with the spelling rules of Old Bulgarian, that is, by the letter **ѭ** (***въpвѭ***, ***въpвѭт***), or, alternatively, by **я** (***въpвя***, ***въpвят***) and even **ꙗ** (***въpвꙗ***, ***въpвꙗт***).

Excerpt (1) below is from a newspaper article published in 1878. It gives a good idea—especially when compared to its normalized version—of some of the key differences between a Drinov type of spelling and the modern orthography (highlights in red and blue):

(1) На телеграммата отъ 10 Юлия Главнокомандующийтъ на войскитѣ
На телеграмата от∅ 10 юли∅ главнокомандващият∅ на войските
позволи изнасяньето на хранитѣ отъ България .
позволи изнасян∅ето на храните от∅ България .
'In a telegram from 10 July, the Commander-in-chief
gave permission to export the food from Bulgaria.'

Except for the Omarchevski model that replaced the two yers **ъ** and **ь** altogether with **ѫ** and dropped silent letters, all other spelling models kept the silent **ъ** and **ь** at the end of the words phonetically ending in a consonant (in this example, ***отъ*** [ot] and ***Главнокомандующийтъ*** [glavnoko'mandujuʃtijt]). Here we have also an example for another redundant letter, **ѣ**, that denoted /ɛ/ in Old Bulgarian (***войскитѣ*** [voj'skite], ***хранитѣ*** [hra'nite]). In Western Bulgarian dialects, the reflex of the vowel /ɛ/ is /e/, while in the majority of the Eastern dialects it is /ja/. After the reform of 1945, a complex rule regulated the replacement of **ѣ** with **е** or **я** depending on prosodic and phonetic factors. The rest of the differences between the two spellings in example (1), are either the result of dialect variation (***изнасяньето*** [iz’nasʲanʲeto] vs. ***изнасянето*** [iz’nasʲaneto]) or introduction of foreign norms—gemination (***телеграммата*** vs. ***телеграмата***) and capitalization of the names of the months and job titles.

Observations on NLIV collections of historical texts show that until 1891, more than a decade after the restoration of the Bulgarian state, different publishing entities followed their own spelling and grammar conventions. That was due to the fact that the elaboration of a fully-fledged language standard or language planning in general were not among the top priorities for the Bulgarian governments after the liberation of the country in 1878 [2]. Cyrillic texts until 1945 contain a myriad of letter symbols such as **ѣ**, **ѧ**, **ѫ**, **ѭ**, etc., which were gradually removed from the modern written language, eventually reducing the number of letters in the alphabet to the current 30. These wide variations of the officially accepted language become a serious hindrance to the success rate of OCR.

3 The Solution: Machine-Readable and Normalized Texts

The goal of the project collaboration is to use the tools developed by the technological partners in CLaDA-BG to minimize and correct errors in the machine-readable texts produced by OCR software, and also to make possible their normalization in order to aid the user, so that s/he would not have to search for a word or expression twice, in the new and in the old spelling. The retrieved search results should include both.

Advancements in the area of accessibility are especially important in the current times, marked by the COVID-19 pandemic. Indeed, as the demand for credible e-resources surges, digital libraries have emerged as vital pathways to high-quality e-books, journals and educational content. Statistics from the world’s leading e-libraries testify to their cultural significance [4].

4 The Approach

4.1 Old Bulgarian Orthography Language Resources

The first major outcome of the work on the project was the preparation and testing of a lexicon of old Bulgarian spelling word forms, to be used for the purpose of assisting OCR. Initially, we decided to opt for a strategy where all word forms from a modern Bulgarian lexicon[3] are transformed to comply with the older orthography [6] developed by the linguist, ethnographer and university professor Stoyan Romanski in 1933. The choice of the prescriptive source was based on its comprehensiveness, the fact that it provides both a detailed and clear definition of the rules and a lexicon. Last but not least, the dictionary of Romanski represents a version of Drinov-Ivanchev orthography that was

[3] The electronic version of [5].

widespread in one of most prolific periods in the history of Bulgarian literature. Many of the literary works created between the two World Wars, are available in modern and Drinov-Ivanchev spelling, which makes much more easier the development of a parallel corpus necessary for the training of a neural network model for normalization. The new "old" version of the lexicon was created using rule-based method in the XML-based CLaRK editor [8] and then imported in FineReader as a user dictionary named CLADABG-MODEL. The testing was conducted in the period March-April 2020. The program ABBYY FineReader (ver. 14 and 15) was used to carry out recognition of 20 pages from issue 1/1881 of the magazine "Наука" ("Science") from the holdings of NLIV, with call number П РЦ-9. All pages are color scanned with an i2S CopyBook A2 scanner at a resolution of 300 ppi, 24-bit, TIFF format, no compression.

4.2 Experiments and Results

The purpose of the test was to determine to what extent the dictionary with old word forms assists the software program in performing OCR of printed Bulgarian texts before the orthographic reform of 1945. The dictionaries used by FineReader are lists of words available in a specific language. The program relies on dictionaries to increase the quality of recognition by reinforcing hypotheses about words included in the dictionary. Custom dictionaries are especially useful in case the text contains many non-common words [1].

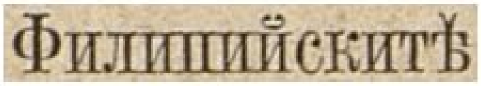
The program has a built-in dictionary only for the modern Bulgarian language. CLADABG-MODEL contains 1,121,872 words from the time before the Fatherland's Front Reform of 1945, including words that are no longer in use or word forms with letters that were gradually removed from the alphabet of modern Bulgarian like ѣ, ѫ, ѧ and so on. Many of the digitized valuable library possessions contain text that is pre-1945, and the purpose of developing CLADABG-MODEL was to test the hypothesis that its use will lead to a higher recognition rate. The amount of the increase, if any, also had to be determined. We used as a main indicator the percentage of misrecognized words[4] in relation to the total number of words. The counting was done manually.

In the course of the test, two other characteristic features of the OCR process and of the software program were measured: the degree of recognition of images in grayscale (as opposed to those in color) and whether and how the FineReader-reported parameter "Low-confidence characters" (expressed in percentage) can serve as an indicator of the success of OCR.

The original paper version of the journal "Наука" is very well preserved, and respectively, the resulting scanned files are close to the optimal characteristics recommended for OCR. However, there is some darkening of the paper, which

[4] Misrecognized are the words in which there is a discrepancy between a letter symbol in the scanned primary word in image form and the same letter symbol in the derivative machine-readable word. It is not considered incorrect recognition if the primary word is spelled incorrectly and the derived word has correctly recognized letter characters, thus duplicating the spelling error.

reduces the contrast and distinctiveness of the letters. Also, the chosen font (widely used back then) makes it difficult for the program to distinguish letters with dominant vertical lines, such as **и**, **п**, **н**, **ш**, **л** (see Fig.1.). The horizontal lines converge, the letters fuse together and further complicate the task for the recognition algorithm.



Филиппийскитѣ

Figure 1. Example of a word with merged letter symbols.

To test the degree of recognition of CLADABG-MODEL, 20 identical pages were scanned, with uniform text and font. The total number of words is 5485 and their average number per page is 274.25.

Minimal training was done, to aid the recognition of traditionally problematic symbols such as **ѫ**, which without prior training always becomes a **ж**.

Table 1. Mean percentage of misrecognized words for 20 color scanned pages, 300 ppi, 24-bit, TIFF format, no compression.

Percentage of misrecognized words		
FineReader built-in dictionary	CLADABG-MODEL	Combined
4,90%	**4,40%**	**4,50%**

A testing was included also for the simultaneous, combined use of two dictionaries, the FineReader built-in Bulgarian dictionary and the CLADABG-MODEL, with recognition performed using two base languages: (1) "Bulgarian" with a standard, present-day set of characters, with the FineReader built-in Bulgarian dictionary, and (2) "Bulgarian before 1945" featuring a character set with added old letter symbols, such as **ѣ**, **ѫ**, **ѧ**, etc., and with the CLADABG-MODEL dictionary. The inclusion of the combined dictionary test was done due to the consideration that when the program works only with CLADABG-MODEL, there is a risk of greater recognition failure in words still in use in modern Bulgarian. The results are summarized in Table 1. The results show that the recognition with CLADABG-MODEL is improved. Although the improvement is not so significant—on average with 0.5% fewer misrecognized words—it shows that this line of research is worth pursuing.

The second test was related to the ability of FineReader to recognize the hyphenation of words split at line-breaks (see Table 2). In case of successful

Table 2. Average number of misrecognized line-break split words per page.

Number of misrecognized line-break		
FineReader built-in dictionary	CLADABG-MODEL	Combined
6,90	**5,55**	**6,05**

recognition, the line breaking is omitted, thus the split words are kept whole, enabling their searching, copying, etc.

The trend of initial slight improvement using CLADABG-MODEL was confirmed by the second test as well. Concerning the recognition difference between color and greyscale pages, the recognition success of the greyscale pages was only slightly better, which does not justify prioritizing the greyscale scanning mode or unnecessary file conversion.

5 A Three-way Collaboration Experience within NLIV and IICT-BAS

The partnership between NLIV and IICT-BAS brought about the intense teamwork between three people—Ivan Kratchanov, a librarian, Laska Laskova, a linguist, and Kiril Simov who is a computer scientist. While the last two shared the same professional physical space in Sofia, the communication with Ivan Kratchanov who is based in Plovdiv, was predominantly via e-mail, chat and video calls. Other factors also played significant role in the development of the project. Neither of the three researchers involved are new to the challenges posed by the interdisciplinary nature of the interaction—Kratchanov, who is Head of the Digital Center at NLIV, has previous experience with digital image processing while Laskova and Simov have worked together for several years on various projects in Natural Language Processing. After the initial discussion of the workflow was concluded with a more or less clear definition of the specific tasks, these tasks were distributed among the three team members with regard to their expertise and access to resources.

The tasks performed at NLIV were related to the selection of digitized materials from different genres, different time periods and different quality of printing, papers, etc. Kratchanov also performed the training and evaluation of the different OCR models. The colleagues at IICT-BAS worked on the creation of lexical resources and their conversion to the different old spelling norms. Another ongoing task for the team members at IICT-BAS is the creation of parallel corpus in several orthography representations.

6 Conclusions

Overall, the benefits of CLADABG-MODEL have been proven and its use is highly recommended. The work on the lexicon will continue in order to streamline the process as a whole, to its efficiency in terms of higher recognition success.

Two are the major reasons for these modest results. In the period from mid-19^{th} century to 1945, many spelling systems were introduced and put to use, while the "old" lexicon represents only one of them, albeit widely accepted, from 1933. One solution to this problem is to create additional "old" versions of the inflectional lexicon that will reflect various spelling models and their codification in monolingual dictionaries, grammars and other documents. Alternatively, we could enrich the "old" lexicon which will encompass several spelling variants for each word form much like a multilingual dictionary. Another reason for the results obtained so far lies in the scarcity of personal names represented in the lexicons, not to mention named entities of other categories, for example organizations or products. We plan to solve this by adding lexical material extracted from manually corrected OCR-ed texts.

Besides training of the OCR software, we envisage to implement a neural network spellchecker for the OCR-ed historical texts. In this case the model will rely on a wider context in order to predict the wrongly recognized words. In order to train the models, we plan to create automatically a parallel corpus with historical and modern texts using the "old" lexicons and pre-trained models.

Acknowledgements

This work was partially supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019.

We would like to thank Petya Osenova for the support during our work and for her comments on the paper.

References

1. ABBYY Technology Portal: Dictionaries and OCR. https://abbyy.technology/en:features:ocr:dictionary_support. Last accessed 8 Sept 2020
2. Andreychin, L.: Iz istoriyata na nasheto ezikovo stroitelstvo [From the History of Our Language Construction]. Darzhavno izdatelstvo "Narodna prosveta", Sofia (1977) [In Bulgarian]
3. Danailova, V.: Basic factors triggering the spelling reform in the Bulgarian Language. Crossing Boundaries in Culture and Communication. **5**(2), 51–56 (2014)
4. Falt, E., Das, P. P.: Digital libraries can ensure continuity as Covid-19 puts brake to academic activity. https://en.unesco.org/news/digital-libraries-can-ensure-continuity-covid-19-puts-brake-academic-activity. Last accessed 11 Sept 2020

5. Popov, R., Simov, K., Vidinska, S.: Rechnik za pravogovor, pravopis, punktuatsiya [Orthoepic, Spelling and Punctuation Dictionary]. Atlantis, Sofia (1998) [In Bulgarian]
6. Romanski, S.: Pravopisen rechnik na balgarskiya ezik s posochvane izgovora i udarenieto na dumite [Orthographic Dictionary of Bulgarian Language with Word Pronunciation and Accent]. Knigoizdatelstvo "Kazanlashka dolina", Sofia (1933) [In Bulgarian]
7. Rusinov, R.: Istoriya na balgarskiya pravopis [A History of Bulgarian Orthography]. Nauka i izkustvo, Sofia (1981) [In Bulgarian]
8. Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A.: CLaRK – an XML-based System for Corpora Development. In: Proceedings of the Corpus Linguistics 2001 Conference, pp. 558–560. UCREL (2001)