Contribution to the Universal Dependencies Treebank of Non-Standard Romanian Texts

Victoria Bobicev[1], Catalina Maranduc[2], Tudor Bumbu[3], Ludmila Malahov[3], Alexandru Colesnicov[3], Svetlana Cojocaru[3]

[1]Technical University of Moldova, [2]"Al. I. Cuza" University, [3]Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova
[1]Chișinău, Moldova, [2]Iași, Romania, [3]Chișinău, Moldova

victoria.bobicev@ia.utm.md, catalinamaranduc@gmail.com, bumbutudor10@gmail.com, lmalahov@gmail.com, acolesnicov@gmx.com, svetlana.cojocaru@math.md

Abstract

Cultural heritage preservation is the one non-transferable duty of any given ethnic or social entity, for it is the essence that defines and identifies each one of them among others. In the specific case of the preservation of culturally significant works of writing, this task includes not only digitizing old books to prevent their loss but also optical character recognition, transliteration of old texts and their annotation. We report our latest contribution to the development and enrichment of a universal dependencies (UD) treebank which contains old texts, regional folklore and other non-standard texts from Moldova and Romania.

Keywords: Cultural heritage preservation, Old Romanian texts, digitizing old books, optical character recognition, transliteration, morpho-syntactic text annotation.

Rezumat

Păstrarea patrimoniului cultural este datoria netransmisibilă a oricărei entități etnice sau sociale, deoarece este esența care o definește și identifică. În cazul specific al conservării operelor literare semnificative din punct de vedere cultural, această sarcină include nu numai digitalizarea cărților vechi pentru a preveni pierderea lor, dar și recunoașterea optică a caracterelor, transliterarea textelor vechi și adnotarea lor. Raportăm contribuția noastră recentă la dezvoltarea Treebank-ului de dependențe universale (UD) care conține texte vechi, folclor regional și alte texte non-standard din Moldova și România.

1. Introduction

Digitisation, preservation and online access to historic literary and cultural treasures are listed among the priorities of the Digital Agenda for Europe. The actions undertaken by the EU include the development of the European Digital Library Europeana[1], supported by the EU Program for Culture. Multiple European research groups and laboratories addressed various problems of creation of linguistic resources by digitisation and recognition of historic and literary heritage (Moruz et al., 2012) through different European projects. Unfortunately, the scientific centres of the Republic of Moldova aren't involved in these actions in spite of their efforts in this domain.

The Government of the Republic of Moldova approved the National Strategy for the development of information society "Digital Moldova 2020" and the Plan of Actions for implementation of this Strategy: the Program "Creation, development and evaluation of the digital content in the RM in 2016 - 2020".

The main aims of the cultural policy for the spaces where the Romanian language is spoken include the study, digitization and preservation of its heritage. The digitization process requires solving a series of problems related to the recognition, editing, transliteration, interpretation and reception of printed Romanian texts in both Latin and Cyrillic alphabets.

Working on these tasks for the Romanian historical linguistic heritage means solving a number of specific problems, namely: the large number of periods in the evolution of the language, the small volume of resources widely distributed, the great diversity of alphabets used in their printing, in particular some Cyrillic-Latin "transition alphabets", the lack of tools for the correct recognition of Cyrillic letters from different historical periods, as well as the lack of lexicons suitable for the period of printing of the resource.

In order to overcome the abovementioned problems a platform has been created which integrates a set of software components for image processing, text recognition and transliteration into modern Latin spelling. It has been adapted for the recognition and transliteration of texts from different historical periods, and for the differences in evolution of the alphabets used for Romanian language printings in Romania and in the present territory of the Republic of Moldova.

2. Our Heritage

We work with old Romanian books in the Cyrillic script.

[1] https://ec.europa.eu/digital-single-market/en/europeana-european-digital-library-all

The researchers of Romania and of The Republic of Moldova have the same problem. The two countries constituted a single state in the past, the historical documents (written in Old Cyrillic) are common, and the regional variants of Romanian spoken in the two countries, with minor differences, are mutually understandable.

The starting point of our work is the scanned text, i.e., the text presented in the form of page images. The sources of these text images are the electronic libraries of texts in this form, e.g., Bucharest Digital Library[2], National Library of Moldova[3].

Table 1 lists the linguistic sources we have been working on recently. The sources are of different historical periods starting with the oldest ones printed in the very first printing houses situated in Moldova in XVII-XVII centuries.

Old Cyrillic fonts, especially of the selected epoch, are much less variable than Latin ones. The usage of the Cyrillic script is connected with the Slavonic liturgical language of the Orthodox Church.

XVII century	Noul Testament, 1648
XVIII century	Fiziognomie, 1785 Ducere către aritmetica, 1785 De obste Gheografie, 1795 Aşezamant, 1786
XIX century	Epistolariu, 1841 Gramatica românească, 1835 Legiuirea Caragea, 1818
XX century	Folclor din părţile codrilor, 1973 Colecţii de reviste 1950-1992

Table 1: Scanned, recognized and transliterated books.

Figure 1 presents small fragments of scanned text of different periods.

3. OCR: problems and solutions

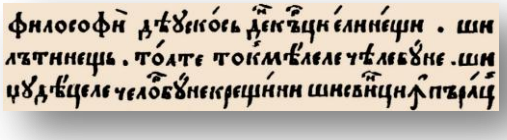
Post-processing of digitized text is a complex task. To solve it, we are developing software that supports expert's efforts in improvement and analysis of the recognized texts. The highest priority task of post-processing is to minimize errors in the recognized text.

The conversion of historical documents from the paper to accessible and searchable electronic form meets two obstacles that are not fully cleared till now. Nowadays state-of-the-art in OCR guarantees relatively good results only on modern texts. For historical typography, results are worse for several causes. Historical fonts vary even in one book, and are less readable. Old paper introduces speckles and distortions. Linguistic components and resources of modern systems don't often know the peculiarities of historical language variations. Each text yields its own specific mix of features and problems, which implies that the quality of OCR for historical documents may vary from perfect to almost unacceptable. The second general problem is produced by the historical orthography and language changes. Most users of digital libraries don't have a good command of old language and desire to use the modern orthography at their search. Any word can have numerous variants in the historical documents because of language evolution and lack of orthography standardization. To get satisfactory replies at search, it is necessary to skip over the gap between modern and new orthography. Availability of texts in original historical orthography differs considerably for different scripts. For example, Romanian Cyrillic script of the 18th century has glyphs that are not supported by most OCR programs.

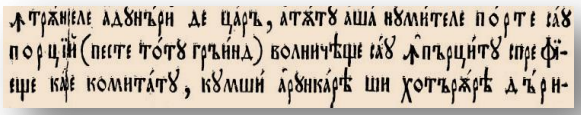
Technology for recognition of the historic and linguistic Romanian heritage printed in the Cyrillic script in the 17th–20th centuries is supported by a pack of the following tools and utilities:

- Alphabets for ABBYY FineReader (AFR).
- Dictionaries (word lists) for AFR.
- Recognition patterns as trained under AFR.
- Selection utility to start AFR with the alphabet, dictionary, and templates corresponding to a specific epoch and location.
- Virtual keyboard.

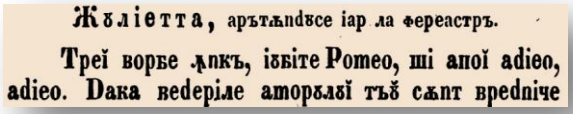
XVII century



XVIII century



XIX century



XX century

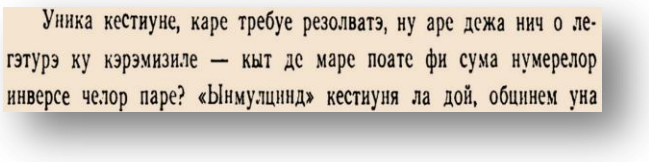


Figure 1: The small fragments of the texts from different centuries.

[2] http://www.bibnat.ro/

[3] http://www.bnrm.md/

The recognition of texts of the 18th-19th century resulted in WER (Word Error Rate) of 3−4.5% and the WER of the 17th century is more than 6%. Figure 2 presents a fragment of Cyrillic text from XVII century after its recognition. In spite of the fact that all letters a clearly seen it is still barely readable by a modern Romanian speaking person due to the specific alphabet which mix Latin, Cyrillic and some Greek-looking letters.

To solve the problem of multiple character sets for the old texts we developed historical alphabets and sets of glyphs recognition templates specific for each epoch. The dictionaries in proper alphabets and orthographies were created in order to minimize the error rate. In addition, virtual keyboards, fonts, transliteration utilities, and other tools were developed for the researchers of old documents.

A special interface is created for the selection of the historical period and the geographical region, where the text was printed. User can choose one of the following variants: Iasi, Bucharest, Târgovişte, Bălgrad (Alba Iulia), Uniev (Cernăuţi), Sas Sebeş, Snagov or Buzău. Within a region the typography should be selected. For example, for Bucharest the system is trained in recognizing the fonts from the Royal Typography and that of the Bucharest Metropolitan Chair.

4. Romanization of Cyrillic

Once the scanned image was processed and the editable and intelligible Cyrillic text was obtained, the transliteration process takes place.

Unusual fonts are difficult for perception even for professionals in linguistics. Therefore, solving the problem of textual cultural heritage dissemination supposes the development of tools for transliteration in common modern Romanian alphabet.

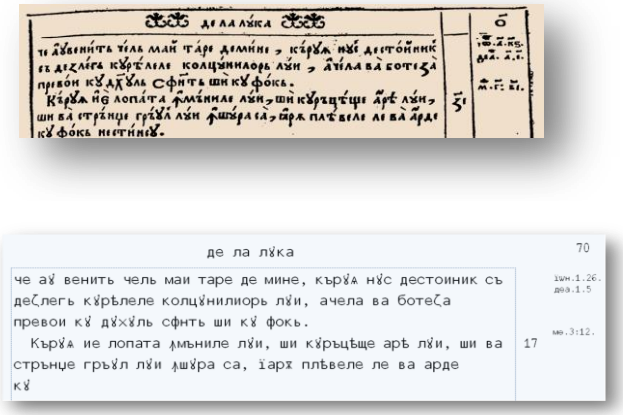


Figure 2: A fragment of a scanned page from New Testament (1648) before and after OCR.

The first problem is presentation of recognized Cyrillic text in computer, especially for transitional and Romanian Cyrillic. In fact, only three fonts in the whole world have old Romanian Cyrillic letters: Kliment STD, Unifont, and Everson Mono only since 2009.

Specific Romanian Cyrillic script of the mid 18th century till 1830 is characterized by two substantial differences from that of the older time. Each period has its specifics and needs specific processing.

Transitional alphabets were used in the Romanian typography since 1830 and until 1860-1870 (Cazimir, 2006). They can be characterized by regular many-to-one mapping of old Romanian Cyrillic letters to the mix of Latin and Cyrillic letters. This mapping could be expanded further to modern Latin Romanian script; slightly different orthography poses an obstacle. The existence of such mapping distinguishes the old Romanian Cyrillic and transitional scripts from Moldavian Cyrillic script that cannot be regularly mapped to the modern Latin script (Ciubotaru et al., 2015). The solution of these problems for the Republic of Moldova faces specific difficulties: the existing resources are scarce and they were printed in diverse alphabets.

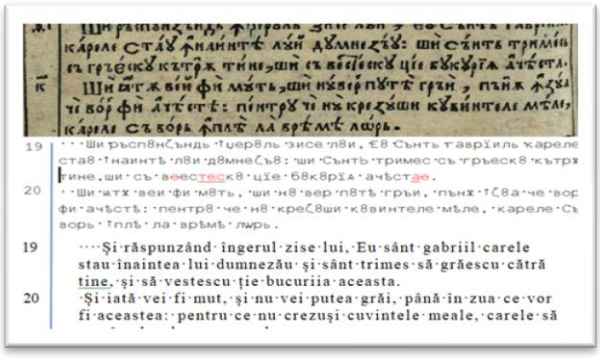


Figure 1: Three forms of an old document: scanned picture, recognized editable Cyrillic text and the text transcribed in Latin letters.

The transliteration of the Moldavian Cyrillic to Modern Romanian Latin was discussed in details in (Boian et al., 2014). The method is rule based; three groups of rules were created manually. Most letters (26 of 31) can be mapped one-to-one as, for example, ш to ş; ц to ţ. Three letters (г, к, ч) can be transformed using context-dependent rules. The letter ы may be transformed in either î or â in accordance with the rule of the Romanian language. The letter я is the most difficult case that can't be fully solved without access to dictionaries. Rules are mostly heuristic and statistical, and more than 20 rules do not cover all cases. This situation exists because MC was not thoroughly designed but is an irregular mapping of Romanian sounds to the Russian letters.

The transliteration algorithm of old and transitional alphabets contains mostly simple rules; the letters in these alphabets are less ambiguous. The old Romanian Cyrillic script reflected the word composition the most accurate. The accuracy of conversion is up to 95% for Moldavian Cyrillic, up to 96% for Transition alphabets and up to 98% for Romanian Cyrillic.

5. Annotation and conversion in UD

The next step after the texts were transliterated was their enrichment with the linguistic information. The texts were automatically processed at UAIC[4] by the Robin-hybrid POS-tagger (Simionescu, 2011), using MULTEXT East project PoS tags (Erjavec, 2004). The set of morpho-syntactic tags for Romanian language developed during the MULTEXT project consisted of 614 tags. This set was quite large with the detailed description of the specifics of Romanian morphology. We simplified the tags keeping 450 tags from the main set and adding around 100 tags to annotate specific elements of the old language. Table 1 contains an example of text enriched by the morpho-syntactic tags.

Old Cyrillic	Modern Romanian	Morpho-syntactic tag
Їwсиф	Iosif	Npmsrn
ф8ци	fugi	Vmis3s
к8	cu	Spsa
ЇС	Iisus	Npmsrn

Table 2: A fragment of text from Noul Testament written in old Cyrillic transliterated in modern Romanian and enriched with morphological tags.

The accuracy of automate morphological tagging was 95 - 96% on various modern Romanian texts; on old texts it was considerably lower. We enriched the dictionary with old Romanian words and manually corrected the annotated old texts and by the bootstrapping method we created the non-standard gold annotated corpus.

Syntactic annotation was obtained by automate annotation by MaltParser and subsequent manual verification and correction by linguists. The convention of annotation is FDG (Functional Dependency Grammar), with labels of classical syntax, with numerous semantic sub-classifications of modifiers. The first texts of our corpus were annotated using the parser trained on UAICRoDepTb (UAIC Romanian Dependency Treebank) and the automate annotation had to be verified and corrected manually as its accuracy on the non-standard old texts was quite poor.

We annotate using dependency grammar formalism developed at UAIC which can be transformed in two formats: the modern syntactic system of Universal Dependencies (UD) with loss of semantic information and into a semantic annotation system by adding information.

The corpus annotated in the initial formalism is registered as UAIC-RoDia DepTB[5] (Romanian Diacronic Dependency TreeBank) and in the UD format is uploaded as a part of UD project Romanian-Nonstandard corpus[6].

The annotated part of the corpus is growing rapidly and has now 15843 sentences and 318869 tokens containing old texts, (1592-1818), and folklore from Romania and Republic of Moldova. The accuracy of automate annotation of the text in this treebank is around 80% and we are working in order to obtain the accuracy over 90% by increasing the gold annotated corpus verified and corrected manually.

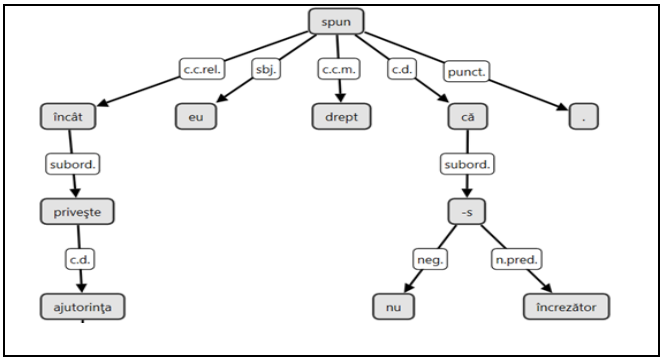


Figure 2: A fragment of a sentence with syntactic annotation opened in a graphic editor for the annotation correction.

6. Conclusion

The Romanian language can be classified as "under-resourced". Under-resourced languages pose important scientific challenges however NLP for under-resourced languages tends to be carried out in isolated and sparse research groups, and the resulting products are often in different formats and standards. We are working on digitizing old Romanian books improving their optical character recognition, transliteration in modern Romanian script and their annotation.

7. Bibliographical References

Boian, E.; Ciubotaru, C.; Cojocaru, S.; Colesnicov, A; Malahov, L. (2014). Digitizarea, recunoaşterea si conservarea patrimoniului cultural-istoric. Akademos, Nr. 1(32), 2014, pp. 61–68.

S.Cazimir (2006). The transitional alphabet. - Bucharest: Humanitas, ISSN 973-50-1401-7. (In Romanian)

C.Ciubotaru, S.Cojocaru, A.Colesnicov, V.Demidov, L.Malahov (2015). Regeneration of Cultural Heritage: Problems Related to Moldavian Cyrillic Alphabet. International Conference "Linguistic Resources and Tools for Processing the Romanian Language". Eds: D.Gîfu, D.Trandabăț, D.Cristea, D.Tufiș. pp. 177-184.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora.In Proc. of the Fourth Intl. Conf. On Language Resources and Evaluation, LREC'2004, ELRA http://nl.ijs.si/ME/Vault/CD/docs/mte-d11f/

Moruz, M.; Iftene, A.; Moruz, A.; Cristea, D. (2012). Semi-automatic alignment of old Romanian words using lexicons. In: Proceedings of the 8th International Conference „Linguistic resources and tools for processing of the Romanian language", Iași, Editura Universității „A.I. Cuza", p. 119–125.

Simionescu, R. (2011). Hybrid POS Tagger. In: Proceedings of "Language Resources and Tools with Industrial Applications", Workshop Eurolan 2011 summer school.

[4] Alexandru Ioan Cuza University, Iași, Romania

[5] ISLRN 156-635-615-024-0

[6] https://universaldependencies.org