Transliteration and Alignment of Parallel Texts from Cyrillic to Latin

Mircea Petic[1,3], Daniela Gîfu[2]
[1]Institute of Mathematics and Computer Science, Chişinău
[2]Alexandru Ioan Cuza University of Iaşi
[3]Alecu Russo Bălți State University
E-mail: mirsha@math.md, daniela.gifu@info.uaic.ro

Abstract

This article describes a methodology of recovering and preservation of old Romanian texts and problems related to their recognition. Our focus is to create a gold corpus for Romanian language (the novella *Sania*), for both alphabets used in Transnistria – Cyrillic and Latin. The resource is available for similar researches. This technology is based on transliteration and semiautomatic alignment of parallel texts at the level of letter/lexem/multiwords. We have analysed every text segment present in this corpus and discovered other conventions of writing at the level of transliteration, academic norms and editorial interventions. These conventions allowed us to elaborate and implement some new heuristics that make a correct automatic transliteration process. Sometimes the words of Latin script are modified in Cyrillic script from semantic reasons (for instance, editor's interpretation). Semantic transliteration is seen as a good practice in introducing multiwords from Cyrillic to Latin. Not only does it preserve how a multiwords sound in the source script, but also enables the translator to modify in the original text (here, choosing the most common sense of an expression). Such a technology could be of interest to lexicographers, but also to specialists in computational linguistics to improve the actual transliteration standards.

Keywords: transliteration, Cyrillic and Latin script, parallel corpus, text recognition, text aligning, digitization.

1. Introduction

In this paper we are interested to present a methodology which helps us to recover Romanian text written with Cyrillic script between 1924 and 1989, that is an adaptation of the Russian Cyrillic alphabet to reproduce the Romanian phonetics by Russian orthographical norms. This alphabet is used till now in Transnistria (Republic of Moldova). In this respect, the process of heritage digitization involves many problems that need to be solved and are related to recognition, editing, interpretation, circulation and reception of printed texts (Boian et al., 2013a). We highlighted aspects of development of main language components: alphabet, lexicon, and orthography, specific for that time. As we move from one period to another, we can use previously elaborated tools and resources (Boian et al., 2011); thus, implementing the principle "from now in the depths of time". We talk to an overview of the Romanian lexicography evolution, from the beginnings to the nowadays, focusing on the monolingual dictionaries. Of course over time some linguistic constructions have changed, but a language should not be judged by these forms, but by its roots, which give meaning to words. They do not change substantially (Haja et al., 2005).

In this order, the aim of this paper is to create a gold parallel corpus of Romanian text written with Cyrillic and Latin script that will help to discover some conventions at the level of transliteration, academic norms (Densușianu, 1894) and editorial interventions.

In our research we accepted the definition for the notion of transliteration as "the process of transcription of a Romanian word from Latin script in its equivalent form written in the Cyrillic script according to the accepted linguistic norms, and vice-versa" (Boian et al., 2013b). The importance of the process of transliteration consists not only in the recovering and preservation of old Romanian texts, but also in the process of informational retrieval from the texts written in the same language, but with a different alphabet and as a subtask in machine translation in the process of transliteration of such words as proper nouns (Deselaers et al., 2009).

The paper is structured as follows: section 2 describes briefly the related work of the historical linguistic resources preservation and texts transliteration from Cyrillic to Latin alphabets; section 3 presents a methodology to create the recognition of old texts and alignment of parallel texts from Cyrillic to Latin script and enriches the standard approach with semantic transliteration; section 4 consists of some obtained results of a group of expert linguists analysis of every parallel letter/lexem/multiwords, and finally section 5 includes some conclusions and directions for further work.

2. Previous Work

Our study combines tools that enable transliteration of the texts from Cyrillic script to Latin script and semi-automatic recovery methods of information (e.g. editor's semantic interpretation) that is used in translation process in order to identify differences of multilingual parallel alignment. For the historical linguistic heritage of Romanian language, the solution of this problem confronts with a specific difficulty – the relatively small number and dispersion of deposited resources. The

difficulties in digitization of this heritage lie in the correct recognition of characters and in the lack of adequate lexicons corresponding to the periods of the texts printing. One of the solutions of the lexicon problem could be aligning of old texts to contemporary linguistic norms (Moruz & al., 2012).

As to OCR of printed and handwritten Cyrillic characters, we can mention a paper (Kornienko & al., 2011) where not only ABBYY FineReader[1] system but also other systems that uses AI techniques, in particular, artificial neural networks. There exists an application of methods based on knowledge technologies to the digital archive and multimedia library for Bulgarian traditional culture and folklore (Pavlov & al., 2011). Problems of transliteration caused by parallel use of two alphabets, Cyrillic and Latin, which appear while processing the written texts in modern Serbian, were solved applying monolingual and multilingual corpora and various e-dictionaries (Vitas & al., 2003). A special application that uses specific resources for the historic period of the printed text is necessary for text verification (Burlaca et al., 2010).

A preliminary study of the method of transliteration for Romanian language was performed by the team at Institute of Mathematics and Computer Science from Chişinău. The authors succeeded to formalize an important number of transcription rules over the standards approved by national authority in Republic of Moldova and Romania. Moreover, they have showed that this process is vague and cannot be fully automated because of the phonology, morphology, and syntax discordance between the linguistic norms of the Romanian language and those accepted in the MSSR (*Moldavian Soviet Socialist Republic*, 1940-1991). The process could be automated partially by formalizing rules of transliteration, manual intervention, and text alignment (Boian et al., 2013b).

3. Methodology

3.1. The corpus

We mention the existence the SR ISO 9:1997 standard which establishes a system of transliteration from Latin to Cyrillic characters, but we have stopped at a corpus from 1955. The evaluations between our standard from this paper and the one from 1997 revealed some differences.
In order to obtain an automated Romanian transliterator for Cyrillic script, we need to test heuristics on a real text of that period. The purpose is to emphasize the differences in writing and pronunciation based on some remarks of similar texts.

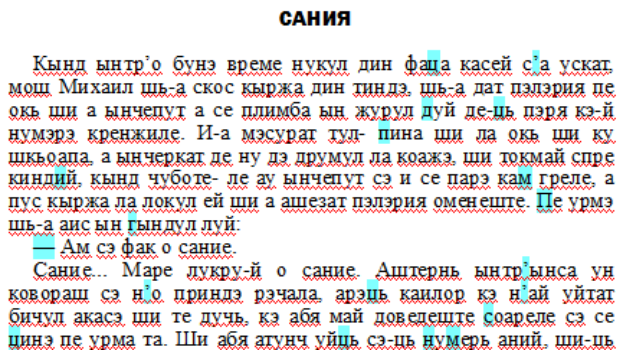
So far, we have identified some transliteration heuristics and tested them on existent Cyrillic-Latin lexicon[2] (Boian et al., 2013b). This corpus was annotated manually at letter/lexeme/ multiword level becoming testing corpus. Once implemented heuristics, we processed the new corpus.

In this research, we have decided to continue with the original Romanian texts processing that are written with the Cyrillic alphabet. The text of the novella Sania (eng. The Sledge) served as a training corpus. It was written in 1955 by Ion Druță and printed originally in Cyrillic script (Fig. 1).

Кынд ынтр'о бунэ време нукул дин фаца касей с'а ускат, мош Михаил шь-а скос кыржа дин тиндэ, шь-а дат пэлэрия пе окь ши а ынчепут а се плимба ын журул луй де-ць пэря кэ-й нумэрэ кренжиле. Й-а мэсурат тулпина ши ла окь ши ку шкьоапа, а ынчеркат де ну дэ друмул ла коажэ, ши токмай спре киндий, кынд чуботеле ау ынчепут сэ и се парэ кам греле, а пус кыржа ла локул ей ши а ашезат пэлэрия оменеште.

Fig. 1 Text in Cyrillic Script

We have started with the process of optical character recognition of the text. We have followed a special previously developed technology of recognition and specialized lexicons. To perform OCR of such texts, it is necessary to train the OCR system to recognize an additional letter ӂ, specific for that period of time. At the OCR time, the letter ӂ not was included in out OCR system. We resolved that by drawing the letter and adding it in available set of characters, **and find 27 words written with this letter in our novella**. Moreover the OCR system lexicon used words from nowadays that is why the OCR process was not accurate.



САНИЯ

Кынд ынтр'о бунэ време нукул дин фаца касей с'а ускат, мош Михаил шь-а скос кыржа дин тиндэ, шь-а дат пэлэрия пе окь ши а ынчепут а се плимба ын журул луй де-ць пэря кэ-й нумэрэ кренжиле. И-а мэсурат тул- пина ши ла окь ши ку шкьоапа, а ынчеркат де ну дэ друмул ла коажэ, ши токмай спре киндий, кынд чуботе- ле ау ынчепут сэ и се парэ кам греле, а пус кыржа ла локул ей ши а ашезат пэлэрия оменеште. Пе урмэ шь-а аис ын гындул луй:

— Ам сэ фак о сание.

Сание... Маре лукру-й о сание. Аштернь ынтр'ынса ун ковораш сэ н'о приндэ рэчала, арэць каилор кэ н'ай уйтат бичул акасэ ши те дучь, кэ абя май доведеште соареле сэ се цинэ пе урма та. Ши абя атунч уйць сэ-ць нумерь аний, ши-ць

Fig. 2 Text in Cyrillic Script OCR-ized

Being trained our OCR system we recognized the text of

[1] http://finereader.abbyy.com/

[2] http://www.math.md/elrr/ - Institute of Mathematics and Computer Science, Moldova, Reusable Resources for Romanian Language Technology, 2003

the novella. The example of the OCR-ized text is presented in Fig. 2. As we see there are several mistakes in the process of optical character recognition. One of the aspects concerns the darker areas that pretend to be some mistakes. In fact only the word луй insteadis not recognized corect as дуй. Also there is the sequence Й-а that is wrong recognized as И-а. Our OCR system being trained, we recognize the text of the novella, (statistically speaking, over 95,8% of words were correctly OCR-ized). Also, the OCR-ized text has been verified and manually corrected.

In such a way, we have obtained the electronic version of Cyrillic script variant of the text. On the other hand, we did the same procedure with Latin script variant of the same text, transliterated manually by expert linguists (Fig. 3).

Cînd într-o bună vreme nucul din faţa casei s-a uscat, moş Mihail şi-a scos cîrja din tindă, şi-a dat pălăria pe ochi şi a început a se plimba în jurul lui de-ţi părea că-i numără crengile. I-a măsurat tulpina şi din ochi, şi cu şchioapa, a încercat de nu dă drumul la coajă şi tocmai spre chindii, cînd ciubotele au început să i se pară cam grele, a pus cîrja la locul ei şi a aşezat pălăria omeneşte.

Fig. 3 Text in Latin Script

3.2. The alignment levels

The obtained corpus allows us to establish an automatic alignment of Cyrillic variant to contemporary Latin variant of the same text at the following alignment levels presented in Table 1:

1	2	3	4
letter to letter	***2877***	***89.6***	сание→sanie
letter to letter group	***164***	***5.0***	пэря→părea
letter group to letter group	***115***	***3.6***	шкьоапа→şchioapa
lexem to lexem	***29***	***0.9***	ынвэлурат→învăluit
word to multiword	***21***	***0.7***	дар→ci pentru că
multiword to multiword	***4***	***0.2***	тоатэ время се гындя → se gîndea
TOTAL	***3210***		

Table 1. Levels of alignment

Legend:

1- Type of alignment level
2- Number of occurences
3- Percent of occurences
4- Example of type

The process was semi-automated, based on the heuristics for transcription of letters and the expert linguists' validation. The corpus[3] is annotated at sentence and word levels, providing morpho-lexical information using UAIC Romanian Part of Speech Tagger[4] (Simionescu, 2011). Below, we have an example from our corpus, for both alphabets Cyrillic and Romanian:

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?><POS_Output>
...
<S id="6" offset="474">
    <W Case="direct" Definiteness="no"
Gender="feminine" LEMMA="mare"
MSD="Afpfsrn" Number="singular"
POS="ADJECTIVE" cyr="Мape" id="6.1"
offset="0">Mare</W>
    <W Case="direct" Definiteness="no"
Gender="masculine" LEMMA="lucru"
MSD="Ncmsrn" Number="singular"
POS="NOUN" Type="common" cyr="лукру"
id="6.2" offset="5">lucru</W>
    <W LEMMA="fi" MSD="Vmip3s"
Mood="indicative" Number="singular"
POS="VERB" Person="third"
Tense="present" Type="predicative"
cyr="-й" id="6.3" offset="10">-i</W>
    <W Case="direct" Gender="feminine"
LEMMA="un" MSD="Tifsr"
Number="singular" POS="ARTICLE"
Type="indefinite" cyr="о" id="6.4"
offset="13">o</W>
    <W Case="direct" Definiteness="no"
Gender="feminine" LEMMA="sanie"
MSD="Ncfsrn" Number="singular"
POS="NOUN" Type="common" cyr="сание"
id="6.5" offset="15">sanie</W>
    <W LEMMA="." MSD="PERIOD" cyr="."
id="6.6" offset="20">.</W>
</S>
```

As a result, we have obtained an important parallel corpus of Romanian text written with Cyrillic and Latin script specific for that period. We have encountered situations in which heuristics have not covered the whole range of the

[3] http://students.info.uaic.ro/~daniela.gifu/LR - Petic Mircea, Daniela Gîfu, Gold Parallel Romanian Latin-Cyrillic Corpus, 2013

[4] http://nlptools.infoiasi.ro/WebPosRo/ - Simionescu Radu, UAIC Romanian Part of Speech Tagger, 2011

text alignment in the process of transliteration Cyrillic alphabet to Latin.

4. Achieved Results

The present corpus consists of 3210 words/13749 characters (Latin script) and 3210 words/13426 characters (Cyrillic script).

Level of convention	Latin script	Cyrillic script	English translation
Transliteration	*şi-a* scos *şi* *şi-ţi vezi* *pălăria* *părea* *şchioapa* *uiţi*	*шь-а скос* *ши* *ши-ць везь* *пэлэрия* *пэря* *шкьоапа* *уйць*	*he took off* *and* *and you see* *the hat* *seemed* *lame* *forgot*
Academic norms	*s-a întors* *deodată* *vreo* *vreun*	*с'а ынторс* *де одатэ* *вре-о* *вре-ун*	*came back* *suddenly* *some*
Editor's semantic intervention	*căutînd a ghici*	*гичинд*	*trying to guess*

Table 2. Examples for Levels of convention

After parallel alignment at **letter/lexem/multiwords** level written in Latin and Cyrillic by automatic means, a group of expert linguists have analysed every word/expression, activity that resulted in the following observations at the level of (Table 2):

A. **Transliteration:**

- Romanian conjunction **şi** is transliterated in Cyrillic in two ways:

a. as ***шь*** in the case the conjunction is followed by a hyphen with an infinitive of a verb;
b. as ***ши*** in the case it is a single conjunction or followed by a hyphen with a pronoun.

- In the case of diphthong we have identified: ***ia → ия***; ***еа → я***.
- The group of letters ***chi*** preceded by the diphthong ***оа*** and placed at the end of the word is transliterated as ***кь***.
- The group of letters ***ţi*** at the end of the word is transliterated as ***ць***.

B. **Academic Norms:**

- By the reform from 1953 the apostrophe from Romanian script is replaced by a hyphen, that is why in the analyzed text we have discovered both variants of scripts;
- There are several lemmas, as article/numeral/pronominal adjective starting with *vre-* (e. g. *Latin vere-unus = Romanian vreun*), which in actual Romanian is written as a single word, but in Cyrillic appears hyphened.

C. **Editorial Interventions:**

- As Cyrillic texts from that period represent an interesting point for our research, there are many interventions of those who transliterated texts at the level of expressions during book editing.

The remarks mentioned above should improve the existent heuristics with new annotation conventions, which increase the degree of accuracy in the process of transliteration of the out test corpus, belonging to the same author representing the same period.

5. Conclusions and Future Work

This gold parallel corpus constitutes an essential support for researchers, and conversions into modern standard text can be used as a support for mastering the developing process of heuristics to recognize alphabets of that specific period. This would allow to recognize words or even expressions and to align texts conforming contemporary linguistic norms. Moreover our corpus can serves as a training corpus for machine learning of transliteration rules.

Development of the proposed technology would provide opportunities to transliterate digitized the Romanian texts from Cyrillic to Latin, customize graphics, offer possibilities for corpora building, and preserve the original texts. It can be used in the building and enrichment of specific linguistic resources with new words/syntagms extracted from digitized resources and certified by expert linguists.

As a next step in the development and implementation of an automated Romanian transliterator for Cyrillic script, we need to validate the remarks on a test corpus collected from more texts of that time, recognized as changing ones.

We have in plan to use the actual methodology for transliterating Cyrillic to Latin, based on the standard of SR ISO 9:1997, for other languages. This standard creates a transliteration system from Latin to Cyrillic, but we have identified few differences. It is necessary to request an updating in order to check many other writings.

For instance, we will use the same methodology for transliterating Arabic to Latin. Actually, we want to compare our results with the standard of SR ISO 233-2:1996. This standard uses the principles of conversion of two writing systems.

Furthermore, we want to investigate the transliteration from Hebraic to Latin. There already exists a standard, SR ISO 259-2:1996.

6. Acknowledgements

We are grateful to our colleagues from the Institute of

Mathematics and Computer Science, Chişinău, Laboratory of Programming System for their contribution in this kind of research.

7. References

Boian, E., Cojocaru, S., Ciubotaru, C., Colesnicov, A., Malahov, L., Petic, M. (2013). Electronic linguistic resources for historical standard Romanian. In: Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language", May 16-17, 2013, Miclăuşeni-Iaşi, România, pp. 35-50.

Boian, E., Cojocaru, S., Ciubotaru, C., Colesnicov, A., Malahov, L., Petic, M. (2013). Language Technology and Resources for cultural and historic heritage digitization . In: Proceedings of the 2nd International Conference on Intelligent Information Systems 2013, August 20-23, 2013, Chişinău, Republic of Moldova, pp. 64-73.

Boian, E., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Malahov, A., Petic, M. (2011). Creation and Development of the Romanian Lexical Resources. In: G. Anghelova, et. al eds, International Conference Recent Advances in Natural Language Processing Proceedings. Hissar, Bulgaria, 12-14 September, 2011, pp. 678-685.

Burlaca, O., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Magariu, G., Malahov, L., Petic, M., Verlan, T. (2010). Applications based on reusable linguistic resources. Multilinguality and interoperability in language processing with emphasis on Romanian, Eds: Tufiş, D., Forăscu, C., Bucureşti, pp. 461–476.

Densuşianu, A. Istoria limbii şi literaturii române. Iaşi, (1894). [Densuşianu, A. History of the Romanian language and literature. Iaşi, – In Romanian.]

http://ru.scribd.com/doc/123035210/Istoria-limbii-si-literaturii-romane

Deselaers, Th., Hasan, S., Bender, O., Ney, H. A Deep (2009). Learning Approach to Machine Transliteration. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30 – 31 March 2009, pp 233–241.

Haja, G., Dănilă, E., Forăscu, C., Aldea B. M. (2005). Dicționarul Limbii Române (DLR) în format electronic. Studii privind achiziționarea. Editura Alfa, Iași.

Корниенко С.И., Айдаров Ю.Р., Гагарина Д.А., Черепанов Ф.М., Ясницкий Л.Н. Программный комплекс для распознавания рукописных и старопечатных текстов. Информационные ресурсы России (2011). №1, с. 35–37. [Kornienko S.I. et al. Program tools for recognition of handwritten and old-printed texts. Informational Resources of Russia, nr. 1, pp. 35-37 [in Russian.]

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

OCR (Optical Character Recognition) Technology

http://www.unescap.org/stat/pop-it/pop-guide/capture_ch01.pdf

Pavlov, R., Bogdanova, G., Paneva-Marinova, D. (2011). Todorov, T., Rangochev, K. Digital archive and multimedia library for Bulgarian traditional culture and folklore. International Journal “Information Theories and Applications”, Vol. 18, Number 3, pp. 276–288.

Simionescu, R (2011). Hybrid POS Tagger. In: Proceedings of “Language Resources and Tools with Industrial Applications” Workshop (Eurolan 2011 Summer School), Cluj-Napoca, Romania, 2011, pp 21-28.

Vitas, D., Krstev, C., Obradović, I., Popović, L., Pavlović-Lažetić, G. (2003). An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts.

http://poincare.matf.bg.ac.rs/~cvetana/biblio/Solun03MATF.pdf