Developing a Technology Allowing (Semi-) Automatic Interpretative Transcription

Daniela Gîfu[1, 2], Mihaela Onofrei[2]

[1] Faculty of Computer Science, University "Alexandru Ioan Cuza" of Iasi
[2] Institute of Computer Science, Romanian Academy - Iasi Branch, Romania
{daniela.gifu,mihaela.onofrei}@iit.academiaromana-is.ro

Abstract. This paper responds to the great interest to humanities researchers who are concerned with the study of the Romanian language in its diachronic evolution: developing a set of tools allowing (semi-)automatic interpretative transcription of scanned Romanian documents written in Cyrillic, in print as well as manuscript forms. The corpus contains old data, belonging to the 19th-20th centuries, in order to develop an automatic recognition and interpretative transcription of Romanian historical newspapers from Cyrillic (Cy) into Latin (La), in both manuscript and printed forms. We think that the present study will have an important impact the humanities research, including that of paleography, history, archaeology and that field of linguistics interested in the study of the language in diachrony, but it will also help the researchers in the field of computational linguistics that develops models for old language, in order to elaborate a diachronic POS tagger so necessary to recover old lemmata.

Keywords: diachronic corpus, transliteration, interpretative transcription, technology for old language analysis, statistics

1 Overview

It is well known that the operation of interpretative transcription of texts written in Cyrillic is extremely laborious, but it will solve a problem of great interest to humanities researchers who are concerned with the study of the Romanian language (including also Bessarabia texts) in its diachronic evolution.

From the perspective of Digital Humanities (history, paleo-linguistics, to name only a few disciplines), this study is innovative because it will open a huge field of research, making feasible automatic indexing and online content-based search in collections of old Romanian documents. The transcripts produced by the machine will be complemented by linguistic annotations, such that the researcher will be able to seize simultaneously: the original Cyrillic-Romanian script, its Latin alphabet transcription, annotated elements as in modern language dictionaries (tokens and lemmas), and even elements of grammar, such as syntactic structures, etc.

The novelty of this study includes two major components: developing a diachronic corpus, called RODICA (*ROmanian DIachonic Corpus with Annotations*)[1], still in its infancy (here, approximatively 4.5 million words) and defining a method to implement a set of tools allowing (semi-)automatic interpretative transcription of scanned Romanian documents written in Cyrillic [1], using the part of our corpus that also contains texts written in Cyrillic and transliterated in Latin (see Table 2), using the transcription rules described in [2]. Note that the team at Institute of Mathematics and Computer Science from Chișinău succeeded to formalize transcription rules over the standards approved by national authority in Republic of Moldova and Romania. [3].

Research will focus on the automatic transliteration of Cyrillic Romanian texts belonging to the 19th-20th centuries, from journalistic genre (written in the Cyrillic alphabet and transliterated in the Latin alphabet). In order to define a methodology to investigate Romanian old language, we consider that the corpus RODICA responds well. It is sufficient to allow an analytical demarche that aims to identify the deviations from the norm that occur in a language, in epochs that are themselves automatically identified statistically.

This study is focused on semi-automatic interpretative transcription according to the way Romanian language written in Cyrillic could be conserved. In this paper, the main objective is the creation of an electronic corpus of old Romanian texts written with the Cyrillic alphabet, from the 19th to the 20th centuries, belonging to journalistic genre in both manuscript and printed types in order to develop an automatic recognition and interpretative transcription of Romanian language tool from Cyrillic into Latin. As training data in the recognition process, we will use 60% of our corpus in Cyrillic alphabet.

The rest of the paper will be organized as follows; in section 2, we mention a few works related to resources and tools related to the analysis of the old language. In section 3, we will state the problem, present our methodology for the developing an automatic recognition and interpretative transcription of Romanian historical heritage writings from Cyrillic into Latin, using the corpus called RODICA. Finally, the paper contains some conclusive statements and suggestions for future work.

2 Previous Work

The future of a language depends on early exposure and on a large number of people who has access to it.

A technology as the one proposed in this project has never been created before for Romanian old texts. Optical character recognition (OCR) has made remarkable progress in the last decade, current systems almost reaching the performance of human readers who are ignorant on the target language. In particular, the character set of the Latin alphabet is recognized at high rates, which decrease in the case of other alphabets (among them – the Cyrillic one [4]), when umlauts appear or when the alphabets are not standard. For Latin scripts, Holley [5] reports accuracy for recognition of printed 19th and early 20th-century characters in the range 81% to 99%.

Recognition becomes much more problematic in the case of handwritten characters, with their quasi-infinite diversity of forms, being the next phase in our research. Especially problem-

[1] http://profs.info.uaic.ro/~daniela.gifu/LR/

atic is the Cyrillic handwriting recognition. Promising results have been obtained recently in recognizing isolated characters and cursives [6]. The major recognition difficulty in the case of continuous writing is the fact that traditional methods require pre-segmentation of data prior to the classification process. For this type of recognition, the best results were obtained using Multidimensional (MD) Long Short-Term Memory (LSTM) type networks [7]. The MD-LSTM networks go through the data set from multiple directions and decide whether, in a meeting point, a symbol should be issued or not. They learn dependencies in a variable length contextual window, which gives them greater flexibility when changing the training data set. The model outlined in [7] implements a multi-layer network that combines recurrent layers with feed-forward layers. A Connectionist Temporal Classification (CTC) type layer makes the decision on the emission of symbols. The major advantage comes from training the network on images and direct transcripts, rendering manual segmentation at letter level superfluous.

Very few results are known about OCR of Romanian printed with Cyrillic. In [8] encouraging results obtained by using an Adobe solution followed by the application of a rule-based transliteration method are reported. As for old Romanian Cyrillic manuscripts, they pose problems even for human readers and fully automatic recognition is an unattained goal so far. This is why we envisage an interactive OCR-ing solution, where the expert is in the loop, playing also a decision-making role. As more fragments of manuscripts will be interpretatively transcribed, they will be used as training data for innovative DL algorithms with the expected result that automatic transcription suggestions become more precise.

3. Methodology

Our method opens a new perspective for the study of our historical heritage, as conveyed by Cyrillic Romanian, in both manuscript and printed form, by using full text search technology. It will enable adaptation of modules that perform linguistic processing: segmentation of the text at the word level (tokenization), morphosyntactic tagging, syntactic parsing, recognition and classification of proper names, disambiguation of word senses, and others. The modern deep learning (DL) technologies, based on neural networks, which allowed us [9] to develop basic language processing tools for more than 50 languages and language variants, will be further refined and enhanced to deal with old Cyrillic Romanian written texts.

An adequate metaphor for this research is a bridge that covers the long way from pixels to content. Indeed, unstructured grouping of pixels in images representing pages of old Romanian-Cyrillic documents will be interpreted and their inner messages deciphered. In order to acquire a collection of digitised Romanian-Cyrillic resources with their corresponding metadata and interpretative transcriptions; organise training sets, we will establish the set of norms (representation formats of intermediary steps in the process of transforming an image of a page (viewed as a sequence of pixels) into a structured display of textual content. Among these representation formats, we mention: the metadata describing a source document and XML representations of the original Cyrillic characters and the final Latin transcribed content (as much as possible, conforming to TEI [10]). These pieces of content refer to titles, running titles and inter-titles, text placed in columns and lines, extra-linear writing (characters inserted above - supra - or under-infra-lines, with the indication of their position with respect to the main elements of the text, marginal additions, literal and Arabic numbers (as for instance, those indicating verses), etc.

After we acquire an important collection of digitized resources (printed, semi-uncial and handwritten) containing Romanian language in Cyrillic writings, covering all historical periods,

of various conditions of quality (noise level, uneven characters, etc.), with and without supra-linear writing, we will add metadata to this corpus, which will provide details about: main language (which must be Romanian), second language(s) (if the document includes words or passages of text in other languages), year of publication, document script (printed, semi-uncial writing, or cursive manuscript), document source (typography), author, level and types of noise (degraded pages, ink stains, creases, dirt, etc.), inclusion or not of supra-linear writing, if there is any critical edition of the text (with indication of source), etc.

Also, in order to build the parallel corpus of original page images in Cyrillic and their interpretative transcriptions in Latin Romanian (UTF-8), we will annotate this corpus with respect to Elements of Content (EoC) in the layout: characters, words, glued words (scriptio continue), lines of writing, paragraphs, supra- and infra-linear writing (words/characters placed above and under lines), border notations, etc. Extract out of this parallel corpus sample images of characters in context, together with equivalent codes, to be used both in training and in evaluation.

To improve the quality of original images, to segment images down to elementary EoCs, to recognize the language different words or sequences of words are written in, to index and search documents based on their EoCs, and to evaluate the recognition processes, we will develop or adapt state-of-the-art visual segmentation software to distinguish EoCs in context. Experiments will be carried out with commercial and open source packages (ABBYY FineReader, for instance). The segmentation software should allow alignment of the EoCs identified in the original digital format of the document and their deciphered textual equivalents. These pointers have a triple role: to link the textual index back into the source document, to support annotations of the expert users in the original document, and to support their corrections related to the interpretative transcription.

We will also train a language recognition system to distinguish among (sequences of) words those belonging to different languages (often used in old Romanian texts: Romanian, Slavonic, Hungarian, Greek, Latin, etc.). This will enable to distinguish foreign words in human translations.

3.1 Romanian Historical Corpus

RODICA is a lexical resource developed based on an important newspapers collection [11, 12], playing a significant aspect in the process of the literary Romanian language modernization, especially in the 19th century, exemplified and analysed in many studies [13, 14, 15]. This corpus structured in four historical regions (Bessarabia, Moldavia, Wallachia, and Transylvania) is statistically described in Table 1.

Importantly, the corpus RODICA represents a first iteration towards building a Romanian Gold corpus, centred on diachronic meta-annotation, and contains over 4.5 million lexical tokens in Latin. The punctuation, the words with less than two characters and the number from the “Total words” have been removed. Note that part of this corpus has been transliterated from Cyrillic to Latin, see Table 2.

Province	Period	Total words	Total unique old words	% (Total old words /Total words)
Bessarabia	1817-2015	643084	53029	8,25
Moldavia	1829-2015	959010	56790	5,92
Wallachia	1829-2015	1372610	67050	4,88
Transylvania	1837-2015	1609230	210180	13,06
Total		**4583934**		

Table 1. **RODICA Statistics in Latin**

Province	Period	Total Chirilic words	% (Total Chirilic words /Total words)
Bessarabia	1817-2015	51084	7,94
Moldavia	1829-2015	18010	1,88
Wallachia	1829-2015	32610	2,38
Transylvania	1837-2015	89230	5,54
Total		**190934**	

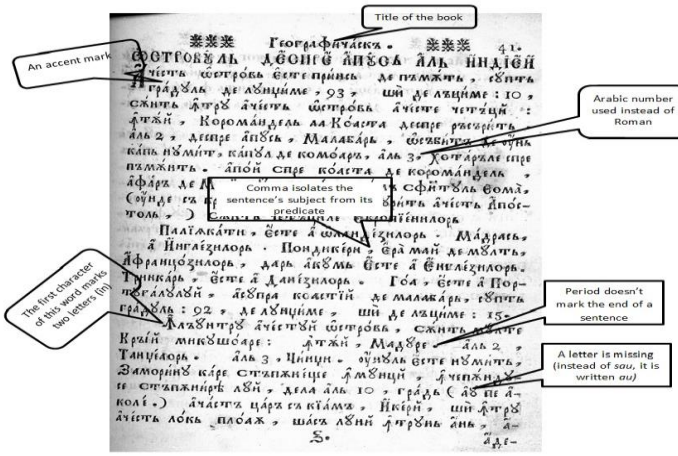
Table 2. RODICA Statistics in Chirilic before transliterated in Latin

3.2 Discussions

For illustration, the Figure 1 contains some examples of unconventional writing. For instance, in Figure 1, it can be observed that period does not always mark the end of a sentence, also, it can be noticed that Arabic numbers are used instead of Romans, that a Cyrillic character must be transcribed in two Latin letters, depending on the letters preceding that character. In addition, the capital letters do not always mark the beginning of a phrase or their own name but are often used without a grammatical explanation. There are also missing letters due to mistakes made by the scribe.

Figure 1: Examples of unconventional writing

Table 3. I **nd Latin**



Cy >1850	Cy <1850	La equiva-lent	Transition alphabet	Phonemes	Letters names
А а	А а	a	А а	/a/	Az
Б б	Б б	b	Б б	/b/	Buke
В в	В в	v	В в	/v/	Vede
Г г	Г г	g, gh	G g	/g/	Glagol
Д д	Д д	d	D d	/d/	Dobru
Є є	Е е	e	Е е	/e/	Est
Ж ж	Ж ж	j	Ж ж	/ʒ/	Juvete
S s	Dz ʤ	dz	Ḑ ḑ	/dz/	Zalu
З з	З з	z	Z z	/z/	Zemle
И и	І і	i	І і	/i/	Ije
Ї ї	І і	i	І і	/i/	Ii
Й й	Ĭ ĭ	i	Ĭ ĭ	/j/	I
К к	К к	c, ch	K k	/k/	Kaku
Л л	Л л	l	L l	/l/	Liude
М м	М м	m	M m	/m/	Mislete
N ɴ	Н н	n	N n	/n/	Naş
О о	О о	o	О о	/o̯/	On
П п	П п	p	П п	/p/	Pokoi
Р р	Р р	r	Р р	/r/	Râță
С с	С с	s	S s	/s/	Slovă
Т т	Т т	t	Т t	/t/	Tferdu
Оу оу	□ □	u	У ȣ	/u/	Uc
Ф ф	Ф ф	f	F f	/f/	Fertă
Х х	Х х	h	Х х	/h/	Heru
Ѿ ѿ	О о	o	О о	/o/	Omega
Ц ц	Ц ц	ț	Ц ц	/ʦ/	Țî
Ч ч	Ч ч	c in ront of e, i	Ч ч	/ʧ/	Cervu
Ш ш	Ш ш	ș	Ш ш	/ʃ/	Șa
Щ щ	Щ щ	șt	Щ щ	/ʃt/	Ștea
Ъ ъ	Ъ ъ	ă, ŭ	Ъ ъ	/ə/	Ier
Ь ь	-	ă, ŭ, ĭ	—	—	Ieri
Ѣ ѣ	Еа еа	ea/e	Ea ea	/æ/	Iati
Ю ю	Ю ю	iu	Iꙋ iꙋ Ĭꙋ ĭꙋ	/ju/	Iu
Ꙗ ꙗ	Ꙗ ꙗ	ia	Ia ia	/ja/	Iaco
Ꙛ ꙛ	Іе іе	ie	Ie ie	/je/	
Ѧ ѧ	ia	ĭa, ea	Ia ia, Ea ea	/ja/, /æ/	Ia
Ѫ ѫ	Ъ ъ	â	Â â	/ɨ/	Ius
Ѯ ѯ	Ѯ ѯ	x	Ks ks	/ks/	Csi
Ѱ ѱ	Пс пс	ps	Пs пs	/ps/	Psi
Ѳ ѳ	Th th	th, ft	T t, Ft ft	/t/ ș aprox. /θ/	Thita
Ѵ ѵ	Ꙋ ꙋ	i, u	І і; У	/i/, /y/,	Ijița

4. Conclusions and Perspectives

We consider that this research responds well both for applicative goals (for enabling effective language chronology analysis using different lexical resources) and for scientific objectives (for exploring the evolution of journalistic language).

Automatic transliteration to the current Latin script and added annotation referring to modern Romanian language are two highly challenging objectives, from both the technological and linguistic points of views, and will open unprecedented research avenues for Romanian scientists and not only.

The success or failure of the study will be estimated according to a combination of the temporal criteria, genre (journalistic) and printed script criteria, as follows: for each historical period of 50 years, a random-per-script sample of 30 pages will be considered.

In the future, we will expand this study for texts belonging to the 16th - 19th centuries, in order to testing the automatic recognition and interpretative transcription of Romanian historical heritage writings from Cyrillic into Latin, in printed as well as manuscript forms.

Acknowledgments

This survey was published with the support of the PN-II-PT-PCCA-2013-4-1878 Partnership PCCA 2013 grant, having as partners „Alexandru Ioan Cuza" University of Iași, SIVECO Romania, and „Ștefan Cel Mare" University of Suceava and of the grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI – UEFISCDI, project number PN-III-P2-2.1-BG-2016-0390, within PNCDI III.

References

1. Onofrei, M., Gifu, D., Bolea, C., 2017. Old Geographical Corpora: a methodology for interpretative transcription at the 9th SpeD 2017, July 6-9, Bucharest, Romania.
2. Petic, M. and Gifu, D. Transliteration and Alignment of Parallel Texts from Cyrillic to Latin. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.), European Language Resources Association (ELRA), 26-31 May 2014, Reykjavik (Iceland), pp. 1819-1823.
3. 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, Chișinău, Republic of Moldova, pp. 64-73.
4. Smith R.W. (2013) History of the Tesseract OCR engine: what worked and what didn't. In Document Recognition and Retrieval XX, edited by R. Zanibbi, B. Coüasnon, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658., doi:10.1117/12.2010051.
5. Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine.

6. Cireșan, D.C., Meier, U., Gambardella L.M., and Schmidhuber, J. (2011). Convolutional Neural Network Committees for Handwritten Character Classification, 11th Conference ICDAR 2011, Beijing, China.
7. Graves, A., & Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In Adv. in Neural Inform. Process. Systems (pp. 545-552).
8. Ciubotaru, C., Cojocaru, S., Colesnicov, A., Demidov, V., and Malahova, L. (2015). Regeneration of Cultural Heritage: Problems Related to Moldavian Cyrillic Alphabet, in Proceedings of the 11th International Conference "Linguistic Resources and Tools for the Romanian Language", Iași, 26-27 Nov., p. 177-184.
9. Boroș, T. and Dumitrescu, Ș.D. (2017). A Convolutional Approach to Multiword Expression Detection Based on Unsupervised Distributed Word Representations and Task-driven Embedding of Lexical Features. In The 18th International Conference on Engineering Applications of Neural Networks (EANN 2017). Athens, Greece, August.
10. Ide, N. Corpus Encoding Standard: Document CES 1, version 1.4, October. http://www.cs. vassar.edu/CES/, 1996
11. Gîfu, D., 2017. Recovering Old Romanian Lemmata, at the 13th International Scientific Conference eLearning and Software for Education, ELSE, Bucharest, April 27-28, 2017. In: Proceedings of eLSE 2017, Ion Roceanu (ed.), Carol I NDU Publishing House.
12. Gîfu, D., 2016. Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language, PhD thesis, "Alexandru Ioan Cuza" University of Iași, Romania.
13. Diaconescu, P. (1974). Elemente de istorie a limbii române literare moderne. Partea I. Probleme de normare a limbii române literare moderne (1830–1880), Bucureşti, pp. 5-6.
14. Andriescu, A., 1979. Limba presei Româneşti în secolul al XIX-lea, Ed. Junimea, Iaşi.
15. Drăgan, I. (1996). Paradigme ale comunicării în masă, Ed. Șansa, București.