HOW TO FIND OUT WHETHER THE ROMANIAN LANGUAGE WAS INFLUENCED BY THE TWO HISTORICAL UNIONS?

DAN CRISTEA$^{1,2}$, DANIELA GÎFU$^{1,2}$, SVETLANA COJOCARU$^{3}$, ALEXANDRU COLESNICOV$^{3}$, LUDMILA MALAHOV$^{3}$, MARIUS POPESCU$^{4}$, MIHAELA ONOFREI$^{2}$, CECILIA BOLEA$^{2}$

$^1$“Alexandru Ioan Cuza” University of Iasi, Faculty of Computer Science, 16 Berthelot St., Iași, Romania
$^2$Institute of Computer Science, Romanian Academy, Iasi Branch, 2 Codrescu St, Iasi, Romania
$^3$Third Institute of Mathematics and Computer Science “Vladimir Andrunachevici”, 5, Academiei St., Chișinău, Republic of Moldavia
$^4$Faculty of Mathematics and Computer Science, University of Bucharest, 14 Academiei St., Bucharest, Romania

{dcristea, daniela.gifu}@info.uaic.ro, {svetlana.cojocaru, kae, mal}@math.md, marius.popescu@fmi.unibuc.ro, {mihaela.onofrei, cecilia.bolea}@iit.academiaromana-is.ro

Abstract

We propose a research methodology intended to prove empirically, using quantitative methods, whether the modern Romanian language was influenced by the two historical union events (1859 and 1918). The changes in language suited for being approached by computational means address: lexicon, morphology, grammatical structure and semantics. Out of these manifestations of language use, we will concentrate in this preparatory study only on the lexicon and, partly, on the semantics. The study is restricted to the written language.

Key words — Romanian language, corpus linguistics, diachronic corpora, similarity.

1. Introduction

It is well known that Romanian people living in the historical provinces, although separated for long periods of time, have conserved a unified language, with rather minor variations in the lexicon, the grammatical structure or the accent. Obviously, the population, speaking rather the same language, facilitated the historical acts of union, because people had the sense of belonging to the same social community. This year celebration of a centenary since the historical act of Great Union was a catalyst for many studies on this matter. But an inverse influence might equally have been manifested, namely the unions themselves should have had an influence on the evolution of the Romanian language. Each union allowed easier circulation of a common press, the administration was the same, and the school functioned under a unique ministry, therefore it must have adopted common language teaching rules.
In this paper we propose a research methodology intended to prove empirically, using quantitative methods, whether the modern Romanian language has been influenced by the two historical union events (1859 and 1918). The changes in language suited for being approached by computational means could address: the lexicon (inventory of lemmas), the morphology (inflectional forms of different part of speeches), the grammatical structure (peculiar syntactic roles that words or expressions could fulfil in language use) and the semantics (senses of words). Out of these manifestations of language use, we will concentrate in this preparatory study only on the lexicon and the semantics, while other aspects were being only vaguely touched. The study is restricted to the written language, since the spoken language cannot be properly sourced.

The second goal of this paper is to initiate a diachronic corpus of the Romanian language (Gifu, 2015), to discuss theoretical means of investigating the distance between different variants of Romanian (Dascalu & Gifu, 2015) in the time proximity of historical union events (this way departing from scholars stating that “no scalar measure can be developed for linguistic distance” (Chiswick and Miller, 2004)), and to support such an enterprise by foreseeing a technology able to deal with the whole way from scanned versions of prints (in both the Cyrillic and the Latin alphabet) up to a corpus containing XML-annotated transcribed texts together with their metadata. In an optimistic perspective, this enterprise will pave the way towards the know-how and a technology able to treat prints and manuscripts in old Cyrillic Romanian. With such an acquisition we will be much closer to a true digitisation of old Romanian writings, meaning by this much more than the mere collecting of scans that could only be turned over page by page, as is the current procedure with most of the old language archives of the languages in Europe or elsewhere (EUROPEANA¹, ITU Library and Archives Service², etc.). Such a perspective would finally allow online access of language researchers into the contents of old prints and manuscripts, which, indexed on multiple criteria, would make possible much better informed forms of research (categorisation of texts, advanced content-based search, comparative studies of syntax and semantics, diachronic and diatopic, diastatic, diaphasic studies on the Romanian language, construction of multilingual parallel corpora, multilingual alignment of Psalters’ verses for Biblical studies, etc.).

2. Methodology of study

The major ingredient of this study should be a representative corpus of extracts of language used in the public area, digitised, brought to an uniform set of UTF-8 (16) characters (in which, for instance, any differences in coding the diachytics would disappear), cleaned of any formatting clues and associated with proper metadata.

The methodology of this study goes through a sequence of operations, starting with the paper sources and ending with statistics that exploit the annotated corpus:

1. establish a balanced (over time and territory) collection of source documents;
2. acquire metadata of primary documents;

¹ www.europeana.eu
HOW TO FIND OUT WHETHER THE ROMANIAN LANGUAGE WAS INFLUENCED BY THE TWO HISTORICAL UNIONS?

3. scan/photocopy the collection of primary documents;
4. deskew and segment the pages down to significant rectangular zones containing grouped chunks of text (titles, columns, lines, cuff notes, interlinear characters, etc.);
5. linearize the 2D display of the chunks in an XML notation;
6. using only a manageable subcollection, interpret the text zones in UTF-8 (16) by an initially trained optical character recognition (OCR) program;
7. repeat in a bootstrapping manner a sequence of correct-retrain-enlarge-OCR operations until the whole collection is correctly interpreted in the original alphabet and the evaluation of the trained OCR shows a flattening to a maximal value;
8. if the alphabet of the original Romanian text is Cyrillic, transliterate the text in context (operation called by linguists interpretative transcription), on a manageable subcollection, the original Cyrillic into the Latin alphabet, by applying an initial set of transliteration norms;
9. repeat in a bootstrapping manner a sequence of correct-retrain-enlarge-transcribe operations until the whole document is correctly transcribed and the evaluation of the trained interpretable transcribers shows a flattening to a maximal value;
10. submit to linguists a subcollection of this collection, uniformly covering the periods and territories under study, for a manual annotation process to tokens, part-of-speech (POS) tags and lemmas;
11. iterating a bootstrapping correct-retrain-enlarge-annotate procedure obtain the annotated form of the collection, tokenized, POS-tagged and lemmatized, until the evaluation of the trained tokenizer+POS-tagger+lemmatizer shows a flattening to a maximal value;
12. once all texts covering the intended period and space are processed this way, organize and index the respective collection; one may keep together all levels of information in the original XML segmentation of pages (from coordinates of corners of textual zones to their Cyrillic and Latin transcription and the corresponding POS+lemmas of words) or place in separate files each specific level; one way or the other, links that would make explicit the connection of the decrypted and annotated text into the original image should always be maintained;
13. extract word embeddings out of the annotated collection;
14. for each act of union, use pre- and post-union data to compute measures and draw graphics;
15. compare and interpret the results (better, with the eye of a philologist).
In the following we will discuss some of these aspects.
2.1. Establishing sources

Our intention is to include in this corpus printed publications covering a period of 30 years before and after each historical union event. The time limits of this corpus would therefore extend to $1859 - 30 = 1829$ up to $1859 + 30 = 1889$, for the first union event, and $1918 - 30 = 1888$, up to $1918 + 26 = 1944$, for the second event. These yields the continuous period $1829 - 1944$ from which our data should be collected (see Figure 1).

![Figure 1: Historical periods around the two union events and the respective provinces involved](image)

To keep the study focused, it is clear that the first 60 years or so of the corpus should collect writings from Moldavia and Wallachia, while for the last approximately 60 years it should contain publications covering all Romanian provinces, including the ex-soviet Moldavia (Bessarabia), which was part of the Big Romania until 1940 and then again between 1941-1944.

2.2. Segmentation and linearization

The zones of printed text on the original scan of a page should be restored for skewing deformations and then segmented at columns, lines, interline characters, markings placed on the cuff of the pages, accents and diacritics, drawings (including capital letters at the beginning of lines), artefacts, etc. The output of this task should be a technology able to produce an XML description of the (rectangular) zones of text that are evidenced on a page, after eliminating the noise, drawings and artefacts. This description will linearize, in an XML specially designed scheme, the zones of text from a 2D page layout (see Figure 2). It should include spatial information about zones/pieces of text in the original image together with their contents. This will contain all links necessary to identify different text spans on the original image. The structure and content of this XML linearization will grow with each of the processing steps applied upon the document. For instance, the Cyrillic decoding of a line of text (after OCR) and its Latin transcription (after transliteration) will be part of this description.

---

3 By coincidence, the year that is considered (Andriescu, 1979) the beginning of the Romanian press, when the two press “veterans” were born: *Albina românească* (Iaşi, ed. Gh. Asachi) and *Curierul românesc* (Bucharest, ed. I.E. Rădulescu).
2.3. **OCR and correction**

In this section we present our experience of using ABBYY Finereader⁴ (AFR) for historical Romanian texts optical recognition. Certainly, AFR is one of the leading instruments of OCR, especially in industry, services, libraries, newspapers, being used in large-scale OCR actions. It performs good preprocessing, especially in analysis of the document structure. The recommended epoch of documents is the 19th century and later. Single glyphs can be trained, but for complete typesets, it is possible to block default recognition patterns and to train glyphs one by one. AFR is proprietary commercial closed software that can’t be modified by users. It is language specific, a “language” consisting of a target alphabet, a word list (lexicon) and a set of recognition patterns. Patterns may be uploaded into AFR, expanded by further training, and finally downloaded in order to be kept. The lexicon is used during recognition to fix ambiguities and to get rid of hyphenation.

The recognition patterns are determined during training and organised in sets of glyphs for specific characters. However, different editions of documents written in the Cyrillic Romanian script impose retraining of the AFR.

The OCR technological process consists of several steps (Malahove et al., 2017). We have used Scan Tailor (in its batch mode) to provide images of the quality suitable for AFR. The program can delete specks, rotate pages, convert images into black-and-white, correct trapezoidal distortions, etc. However, most of AFR corrections should be performed manually, one by one. Moreover, the page structure cannot be restored automatically, especially for historical books that contain many page decorations. A specific flaw arose with some old texts in Romanian Cyrillic that have many overline signs and interline letters. Sometimes AFR splits one line containing overline letters and signs in two. A roundabout solution is to formally increase the image density. This does not imply re-scanning but is equivalent to decreasing the calculated image size. By default, AFR stores the results of OCR as MS Word files.

⁴ [https://finereader.en.softonic.com/](https://finereader.en.softonic.com/)
One of the actions following OCR is replenishing word lists. To work with word lists we have chosen Notepad++, a text editor that can process Unicode and provides search and replace using regular expressions. TextFX extension permits to sort word lists.

Several glyphs belonging to Cyrillic Romanian were not covered by Unicode until the introduction of Cyrillic Extended-B block [A640-A69F] in 2009. Since then, the following fonts appeared that cover this Unicode block. Unifont is a bitmap font and can be used only for preview. The single monospace font for text editors is Everson Mono but it looks ugly. Several other fonts are Kliment STD, Monomakh Unicode TT, Ponomar Unicode TT, Menaion Unicode TT, Segoe UI, and Segoe WP. Calibri and Microsoft Sans Serif cover Cyrillic Extended-B Unicode block only starting with Windows 10. Virtual on-screen keyboard can be used to enter these glyphs, and BabelMap to visualize Unicode fonts.

Most often, OCR software introduces errors. Manual correction is possible even in AFR before storing the OCR result, but exporting it in a more comfortable environment, will ease the correction operations. Improving the OCR result can be done in two ways: by comparing the output against a word list and by using automatic correction rules. Word list creation is a complicated problem because electronic dictionaries of old word forms do not exist. Possible sources are: scientific editions of old texts with manual conversion in the modern script, or modern dictionaries transliterated to the old Cyrillic Romanian script.

Iterative OCR correction using the recognized text itself could also be an option. Each word can be isolated on a separate line, hyphenations deleted, and the result sorted after deleting repeated words. The resulting word list can be checked manually, and then uploaded into AFR to repeat OCR with supposedly better quality. In such a manner, the word list of the corresponding epoch can be expanded with each new text.

Automatic corrections rules can be deduced by comparing the OCR output (Test) against the same span of text manually revised by experts (Gold). These revisions should be made in close resemblance with the original text and not with the modern language. For instance, the original can miss vocals in writing the name of God (Dnmul instead of Domnul) and, very often, cratima is not placed between a clitic and an auxiliary (iau dat instead of i-au dat) or between a negation and a clitic (nu l caută instead of nul caută).

### 2.4. Transliteration and correction

The most needed post-OCR processing is transliteration of the text into the Latin script (Colesnicov et al., 2014). This can be done in correspondence with more scripts covering the period 1812-1989: the Cyrillic Romanian script (CR), the Transitional Romanian script (TR), and the (Soviet) Moldavian Cyrillic script (MC). We will also refer to the Modern Romanian Latin script (since 1904) as to MRL.

There are regular dependences that support conversions CR → MRL and TR → MRL. Most letters are mapped one-to-one, for example, φ → f, but several mappings are context dependent. Examples are: r → gh (before e, u, i); r → g (otherwise); ↑ → i
How to Find Out Whether the Romanian Language Was Influenced by the Two Historical Unions?

(_before m, n, m, n; ↑ → îm (before b, n, b, p); ↑ → în (otherwise). In some texts, the letter ↑ (în) is used differently.

These rules are valid both for CR and for TR, because only the font shape differs (Uncial or Antiqua) but Unicode codes are the same. Transliteration MC → MRL is more difficult because MC was developed following a political decision without any scientific background. It is an irregular representation of Romanian sounds with Russian letters. One-to-one mapping of most letters and context driven rules exist; examples are previous rules for ф and г.

Colesnicov et al. (2014) report maximum one error per page, which approximates to 0.15% of erroneous words. In a juridical text, for instance, all errors were of the type я → ia (instead of the correct ea), e.g., примесь → (primiască) → primească. One possible solution to this would be to generate both forms of the target word and let a dictionary decide. In our opinion, the issue of transliteration should not hide language differences. For instance, the following forms should be maintained: Loran instead of the modern Laurent (a foreign proper noun), from Лоран; dizain instead of design, from дизайн; câne instead of câine, from къне; eşire instead of ieşire, from ешuire.

2.5. What and how to compare

We will suggest in this section a distance measure between two corpora, which, as we will see, could be easily generalized, and out of which the question put in the title to be answered.

Lexical distance. If we note by: $l_0$ the lexis of the corpus $c_0$ and $l_1$ the lexis of the corpus $c_1$, then we can write: $l_1 = l_0 + p - n$, where: $p$ (the positive lexis) is the lexis appearing in $l_1$ and not appearing in $l_0$; $n$ (the negative lexis) is the lexis appearing in $l_0$ and not appearing in $l_1$. The +/- operations are reunions and differences of sets (Figure 5).

Figure 5: Comparing the lexis of two corpora

The lexical distance (noted here \( \mathcal{L} \)) between corpora $c_1$ and $c_0$ should therefore be: $l_1 \setminus l_0 = p + n$. A normalised distance could be obtained by: $D_{l_1/0} = ([p] + |n|)/([p] + |c| + |n|)$, where $c$ is the common corpus (intersection) and $|x|$ is the dimension of the lexis $x$. This way, the distance is 0 if both $p$ and $n$ are null, and it is 1 if the common corpus $c$ is null.

It can be easily proved that this function is indeed a distance, i.e. it has the properties:

a). $D(x,y) \geq 0$;

a') $D(x,y) = 0$ \( \Rightarrow \) $x = y$;

b). $D(x,y) = D(y,x)$;

c). $D(x,z) \leq D(x,y) + D(y,z)$, for any $x, y, z$. 
Using the notations in Figure 6, the most obvious inequality that should be proven is $d_3 < d_0$, i.e. after a union event it is normal to suppose that the distance between languages in the provinces decreased. If, however, an inequality such as $d_4 < d_5$ is shown, it signifies that the language of Moldavia (M) is more stable than that of Wallachia (W). On the contrary, if $d_3 < d_4$, the interpretation is that, in evolution, the language of W is more stable than the one in M. An inequality of this kind would place the province showing the minimum change at the core of the evolution of Romanian language. A comparison can be made with Modern Italian, which coalesced on the core of the Tuscan dialect (Dante's language).

Other interesting inequalities:

- $d_3 < d_1$ => after union the language in W became more closed to the one in M than it was before the union;
- $d_0 > d_2$ => the language in W was more distant to the one in M before the union than after it;
- $d_3 < d_2$ => after union the language in M became more closed to the one in W than it was before the union;
- $d_0 > d_1$ => the language in M was more distant to the one in W before the union than after it.

Variations of this distance can be suggested by placing frequency filters in the formation of sets $l_0$ and $l_1$. For example, $l_i^f$ would be sets consisting only of frequency lemma equal to or higher than 2 (no hapax legomena). In notations of this kind, we would have: $D^f_1, D^f_2, ...$ distances between two languages on increasing frequency criteria. We expect that differences between distances will flatten the more frequency filters are applied, because fewer words are taken into account, ie: $D^{k+1}_m - D^k_m < D^{m+1}_m - D^m_m$, for $m > k$.

2.6. Semantics (senses of words)

Word embeddings show promise as a diachronic tool. In (Hamilton et al., 2016) a robust methodology was developed for quantifying semantic change by evaluating word embeddings against known historical changes (events). We will follow a similar approach:

- Separate word embeddings, for each Romanian province and each historical period (see Figure 1), should be created, by using state of the art techniques, like word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).
HOW TO FIND OUT WHETHER THE ROMANIAN LANGUAGE WAS INFLUENCED BY THE TWO HISTORICAL UNIONS?

- Having these word embeddings, the semantic distance between written languages in different provinces in the temporal vicinities of a historical event can be measured, and a similar analysis as in the case of the lexicon (see Figure 7) can be performed.

In order to compare word vectors from different embeddings we must ensure that the vectors are aligned to the same coordinate axes because low-dimensional embeddings will not be naturally aligned due to the stochastic nature of word2vec and GloVe. In particular, both these methods may result in arbitrary orthogonal transformations, which do not affect pairwise cosine similarities between words from the same embedding but will preclude comparison of the same word across different embeddings.

Let define \( W^{(t,p)} \in \mathbb{R}^{d \times |V|} \) (\( d \) being the dimension of embedding and \( |V| \) the dimension of vocabulary) as the matrix of word embeddings learned from the corpus of a period \( t \) and a province \( p \). We will align two embeddings with the same vocabulary \( V \), while preserving cosine similarities, by optimizing:

\[
R = \arg\min_{Q} \| Q W^{(t_1,p_1)} - Q W^{(t_2,p_2)} \|_F
\]

with \( R \in \mathbb{R}^{d \times d} \). The solution corresponds to the best rotational alignment and the minimum itself \( \| Q W^{(t_1,p_1)} - Q W^{(t_2,p_2)} \|_F \) can be viewed as a measure of the semantic distance between the language from \((t_1,p_1)\) and the language from \((t_2,p_2)\). This distance can be used to pursue a similar analysis as in the case of the lexicon.

The aligned embeddings can also be used for a more refined analysis at the word level. For a particular word \( w \), the cosine similarities between the two aligned embeddings of \( w \) can be used as the semantic similarity of the word \( w \) in the language from \((t_1,p_1)\) and the word \( w \) in the language of from \((t_2,p_2)\). Looking at the nearest neighbours of \( w \) in the two aligned embeddings can give an indication about how \( w \) was used in \((t_1,p_1)\) and in \((t_2,p_1)\). As an example of how this can be visualized see Figure 7, but, apart of time, a province dimension will be added.

Figure 7: Visualizing changes in word meaning (from (Hamilton et al., 2016))

2.7. Other distances

Etymological distances. If we apply etymological filters to the lexicon of a corpus, we can highlight \( l_{LAT}^i \), \( l_{SL}^i \) for the Latin lexis, Slavic lexis etc. of a corpus \( i \) and, hence, the corresponding distances: \( D_{LAT}^{ij} \), \( D_{SL}^{ij} \), etc. (see Section 2.5). It would be interesting to reveal and interpret increases or decreases of these distances in time.
eDTLR (Cristea et al., 2011), with its chronological sources of citations and etymology marked, could be a good resource for applying etymological filters.

**Morphological distances.** Measuring morphological distances between variants of one language means counting differences in word formation (derivation and compounding) and in flexional rules (paradigms, as in conjugation of verbs and declension of nouns and adjectives), for common lexicals. Examples of this kind are: number of modified paradigms, number of differences in endings, migration of words from one paradigm to another, etc.

**Syntactic distances.** Syntactic distances between languages have been intensively studied in Contrastive Linguistics and Foreign Language Learning. Of interest to our study would be to compare differences in syntactic structures between variants of language, as for instance modification of frequencies of occurrence of certain syntactic roles. In order to do that, the corpus should be annotated by syntactic parsers trained on different regions and periods of time.

### 3. Discussions and conclusions

Languages have been studied in comparison for a long time and there is a vast literature on the issue, which, for lack of space, we will not browse in this study. One point of comparison is their historical roots: “The main metaphor that is used to explain the historical relationship is that of the language family or family tree.” (Crystal 1987, p. 292). The main difficulty that has fuelled the pessimistic vision regarding the computation of a linguistic distance (mentioned in the Introduction) comes from “the myriad characteristics that makes up the structure of languages” (Chiswick and Miller, 2004). In this paper we have tried to overpass this barrier in two directions, the lexicon and the semantic content of words and to suggest other possible distances, with the goal of fulfilling the vivid curiosity of how much the two unions in the recent history of Romania have influenced the formation of the modern language.

We are aware of the dangers to misinterpret the results of such a study. Wrong conclusions could come from more directions, but the most pregnant ones are the insufficiency of digitised resources and the unbalanced quantity of textual data in different periods and provinces. Moreover, OCR and lemmatization errors could also induce wrong arguments. The process of manual correction of a corpus of the required dimension being extremely costly, a study of this kind will certainly be feasible in synchronicity with the acquisition of a diachronic corpus of old Romanian language.

The methodology of study for comparing variants of language discussed in this study could suggest more sophisticated solutions applicable to collections of spoken Romanian language and covering modern and contemporary periods. In an enterprise of this kind, the Portal of spoken Romanian and the speech processing technologies that make the core of the ReTeRom project (acknowledged below) will certainly be of great help.
HOW TO FIND OUT WHETHER THE ROMANIAN LANGUAGE WAS INFLUENCED BY THE TWO HISTORICAL UNIONS?

Acknowledgements

This study was realised as part of the research theme of the Natural Language Processing group from the Institute of Computer Science of the Iași branch of the Romanian Academy and was partially supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 / 73PCCDI (ReTeRom), within PNCDI III and by the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286. We are grateful to the libraries of the Romanian Academy in Bucharest and Iași and the Central Universitary Library in Cluj, which generously have allowed us the access to many resources in electronic form. Other XIXth century resources have been offered by Gabriela Haja, from the “Alexandru Philippide” Institute of Philology and Literary History in Iași, Cătălina Mărănduc from the Faculty of Computer Science of the “Alexandru Ioan Cuza” University of Iași and the Natural Language Processing group from the Institute of Mathematics and Computer Science in Chișinău, to whom we address our thanks.

References


---

6 https://profs.info.uaic.ro/~daniela.gifu/PhD%20Daniela%20Gifu%202016/PhD%20thesis%20Daniela%20Gifu%20final.pdf

7 http://dl.acm.org/citation.cfm?id=3078093