Diachronic Evaluation of Newspapers Language between Different Idioms

Daniela Gîfu

Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, Romania
daniela.gifu@info.uaic.ro

Abstract

Due to various reasons, it is not rare that two cognate languages become strained for a period of time, only to become closer for another period of time. Traditionally the degree of similarity was assets by linguistics on the basis of their expertise. However, it is hardly possible to cover a large material only by human effort. We present a methodology of diachronic investigation on news corpora which determines the degree of similarity between cognate languages.

1 Introduction

The present work investigates the linguistic crisis that affects the journalistic laguage in two countries, Romania (including three historical regions: Moldavua, Tranylvania and Wallachia), and the Republic of Moldavia (known as Bessarabia), which until the early 19th century were one state. This linguistic contrastive study between Romania and Bessarabia allows intercepting many similarities, especially in diachrony. The similarities of the Romance languages are becoming more numerous, as we descend deeper into past. [Densuianu, 1902]. Other important differences were also detected, perhaps due to the influence of Russian language reflected on the Bessarabian language, starting from the middle of the 19th century. It is also important to note that starting with the 19th century the Romanian language was influenced for more than 30% by French and Italian (two Romance languages as Romanian). We analyse, via automatic corpus methodology, the similarity of the two languages, between two periods – before the Second World War and after the fall of communist regime.

The methodology we present is language independent and it can be applied to any two corpora, let's call them target and source. In a nutshell, we first determine the characteristics of each of the four corpora and then we compute the similarity of pairs extracted from target and source corpora, on the basis of these characteristics. We take into account all levels of linguistics analysis in order to derive the language characteristics of a language: lexical, morphological, syntactical, semantically and discourse level respectively. We use a large suite of statistical methods in order to determine.

The similarity considering both words, via word embedding techniques and topics, via LDA type analysis. The methodology we present is offers a basis for future large-scale studies, having a large impact on reducing the amount of human effort required by socio-historical linguistic analysis of language idioms in general.

The results of this contrastive analysis highlight the significant changes in the distribution of terms that best reflects the differences in writing style, ranging from sentence and paragraph structure, to topic cohesion. Finally, a formula computes the similarity in a complete and objective way.

In order to meaningfully carry out this analysis we compiled a corpus of journal articles from the geo-political distinct cognates: Romanian and Bessarabian. A large corpus (over 2.6 million lexical tokens), chronologically ordered since the second decade of the 19th century (1817-2015), was developed, structured in four independent collections of publications corresponding to Moldavia – 68373 words, Wallachia – 143612, Transylvania – 2294108 words, and Bessarabia – 92499 words. Based on this corpus we explore the diachronic phenomenon in order to identify statistically Romanian epochs reflected on the printing press and linguistic similarities from Bessarabian press. The Republic of Moldavia was a part of Romania (including three Moldavia, Wallachia, Transylvania) until 1812, and then from 1918 to 1941, becoming an independent state after 1991.

These texts can form the basis of an analytic process that aims to capture the semi-automatic deviations from the current norm. The automatically investigation offers a solution for historian as well, and historical significant correlation in the word usage may be discovered. In fact, diachronic analysis of cognate languages provides clues and insights into what the society considered adequate responses to social problems at a given moment. The rest of the paper is organized as follow: section 2 presents a brief review of relevant literature, section 3 depicts the corpora in details and the methodology, section 4 describes the analyse and interprets the results. Finally, the survey conclusions and future work are given in section 5.

2 Related Work

Many previous works [Leech et al., 2009: Davies, 2013] have focused mainly on the linguistic interpretation of the statistically results. Their hypotheses were based on the ways language chnages without considering their causes.

It has been established that some genetically related languages have a high degree of similarity to each other [Gooskens, 2006; Gooskens et al., 2008]. Various aspects present relevance when investigating the level of relatedness between languages, for example orthographic, phonetic, syntactic and semantic differences. The phonetic alterations have an orthographic correspondent, thus an alphabetic character correspondences [Delmestri and Cristianini, 2010].

The diachronically comparative studies of the Romance languages expose the presence of many similarities [Densuianu, 1902]. Latin language, the origin of Romanian, French, Italian, Portuguese, Spanish, was the starting point, but issues about substratum, superstratum and adstratum which contributed to differentiate languages were not set aside.

The development and use of software for natural language processing (NLP) highlight the defining aspects of the Romanian printing press (morphological and syntactic analysis, semantic analysis and, more recently, pragmatic analysis) that have many similarities to that of Bessarabia on the time axis that we have chosen. The rich literature tells its own story regarding the usefulness of technology and information services [Carstensen et al., 2009; Jurafsky & Martin, 2009; Manning & Schütze, 1999; Cole et al., 1998; Tufiș & Filip, 2002; Cristea & Butnariu, 2004; Trandabăț et al., 2012, Popescu & Strapparava, 2013, 2014, Gîfu, 2015].

Until now, the Romanian diachronic phenomenon was analysed using various methods. One of them relies on the comparison of writing styles according to various indices: text features [Gîfu et al., 2016], textual formality [Eggins and Martin, 1997], and textual styles [Biber, 1987]. Another one is based on machine learning approach to explore the patterns that govern the lexical differences between two lexicons [Gîfu & Simionescu, 2016].

3 Corpus

A large corpus (over 2.6 millions lexical tokens and 6500 pages), chronologically ordered, since the second decade of the 19th century, was developed, structured in four independent collections of publications corresponding to Moldavia (68373 lexical tokens), Wallachia (143612 lexical tokens), Transylvania (2294108 lexical tokens), and Bessarabia (92499 lexical tokens) (see Table 1 for descriptive statistics).

Nowadays the first three regions form Romania, and Bessarabia was a part of Romania until 1812 and then from 1918 to 1941, becoming an independent state after 1991.

Region	Period	Total lexical tokens	Sources
Bessarabia	1817-2015	92499	Basarabia reînoită; Curierul; Candela; Deșteptarea; Viața economică din Bălți; Solidaritatea; Ehos; Buletinul Arhiepiscopiei Chișinăului; Cuvânt moldovenesc; Ardealul; Basarabia; România nouă; Sfatul țării; Democratul Basarabiei; Glasul Basarabiei; Luminătorul; Dreptatea; Basarabia Chișinăului; Literatura și artă; Moldova Socialistă; Jurnal; Contrafort; Jurnal de Chișinău; Moldova suverană; Ziarul de gardă.
Moldavia	1829-2015	68373	Albina românească; Convorbiri literare; Curierul. Foaia intereselor generale; Constitutionalul; Moldova Socialistă; Scânteia; Noutatea; Deșteptarea; Bună ziua, Iași; Ziarul de Vrancea; Monitorul de Vaslui; Evenimentul regional al Moldovei; Imparțial.
Transylvania	1829-2015	2294108	Organulu Luminarei; Gazeta de Transilvania; Gazeta Transilvaniei; Telegrafulu Românu; Foaia pentru Minte Anima și Literatură; Telegraful român; Transilvania; Federațiunea; Gura Satului; Albina; Telegraful Românu; Familia; Aradu; Patria; Chemarea tinerimei române; Dreptatea; Aradul; Curierul creștin; Vatra

			românească; Echinox; Adevărul de Cluj; Făclia; Monitorul de Cluj; Bihoreanul.
Wallachia	1847-2015	43612	Curier românesc; Buletin. Gazeta oficială; România; Curierul românesc; Pressa, România liberă; Românulu; Timpul; Literatorul; Albina; Deșteptarea. Foaie pentru popor; Adeverul; Curierul artelor; Dimineața; Universul; Viitorul; Curentul; Universul literar; Adevărul; Adevărul literar și artistic; Scânteia; Romania literară; Dimineața copiilor; Evenimentul zilei; Gândul; Ziua; Ziua news; Ziua veche;

Table 1. General corpus statistics

In other words, we talk about four Romanian idioms, covering two linguistic registers (journalistic, literature). To each text the following identification information are assigned (regions, year, publication, author).

It is also important that this corpus represents a first iteration towards building a Gold corpus for each region, centered on diachronic meta-annotation. It was prepared during 2 years. First, the corpus was edited in PDF, so we applied the boiling-plate technology to obtain raw text in TXT format (UTF-8 encoding), using Java PDF Library - Apache PDFBox. Then several corrections were made on the raw texts. Second, the processing phase continues with: segmentation, tokenization, lemmatization, part-of-speech, and NotInDict Markup using the UAIC POS-Tagger [Simionescu, 2011].

The result of the processing stage is an XML file that will be forwarded for other data processing. Moreover, we apply GGS grammar rules over the previous file. The GGS rules practically help to the disambiguation of the hyphen. In other words, one can understand when it is about hyphenation at the end of a row and when it deals with the components of the structure of certain words.

4 Methodology

We build diachronic vectors from corpus for each word, keeping on each slot the number of occurrences for a specific year. There are two variants of these vectors that we build, depending on whether different ortho-lexical realizations of the same word are considered the same, thus they count as one vector, or they lead to distinct vectors. The lexical vectors are relevant in time classification tasks, but less useful for topic identification. Consequently, we use one or the other set depending on the task that we need to resolve.

A snap-shot from a typical vector looks like:

```
768 pace /(EN) peace 1 1865 1 1868 17
1877 15 1878 3 1880 1 1897 4 1900
```

768 represents the total number of occurrences in the whole corpus, “pace”, Romania for peace, is the word and the occurrences of this word precedes the year. In this particular case, is easy to spot a variation in the period of 1877 and 1878, which, not incidentally, corresponds to an independence war fought exactly in those years. These types of non-random variances represent the basis for a diachronic analysis. In fact, each epoch is determined by a certain distribution of words.

As some topics of interest change over the time, the distribution of words in newspaper reflects this phenomenon accurately. Thus, by employing a suite of statistical test we can determine no-random changes in the word distribution. In [Popescu & Strapparava, 2013, 2014] was showed that there are a short period of few years within each many words change their distribution. As such, this specific period represents a transitional buffer between epochs. To determine the buffer period we apply to the from year to year. In particular we used three non parametric tests: Welch, run and ratio test.

We test respectively whether two samples come from the same statistical population, or whether there is a large variance with respect to the mean, or the ratio of change from year to year shown an upward or a downward trend.

For a very large corpus, like Google books for example, one can chose an arbitrary set of topics to investigate, but in this case we have a limited amount of data. Thus, we need first to indentify the topics that are represented in our corpus. For this we apply the LDA algorithm. At this step we use the non photo-lexical vectors are used. We filtered out set 25 topics the following topics for the target corpus, i.e. Romanian, like:

```
război, literatură, partide, stat,
pământ, muncitor, artă, sat, partidă /
(EN) war, literature, parties, state,
land, worker, art, village, party
```

For these topics the following epochs have been identified:

1832-1856	1920-1940
1856-1877	1940-1980
1877-1912	1980-1990
1912-1920	1990-2015

Table 2. Romanian Epochs in Newspapers

Considering this epochs as categories we build an SVM classificatory over whole target corpus (Weka implementation). We classified each news from the source corpus, i.e. Bessarabian. First thing we wanted to check was whether the classification is able to pin point correctly the source news. This will give a fairly accurate indication whether there is indeed a similarity over the epochs between the two cognate languages, or the model will assign a more or less random epoch to the source news. We obtain an accuracy on epoch prediction of almost 78%. This figure indicates that the classificatory works correctly on the source corpus.

We averaged over the classificatory confidence for each epoch separately. We take this parameter as indicator of the similarities between the cognate languages, because, one we know that the classificatory is appropriate, the confidence reflects the similarity. In Table 2 we present the figure for each epoch separately.

As Table 3 shows, the similarity varies over epochs. While these figures are not a direct measure of the similarity of the languages, they represent an objective indication of the high and very high overlapping between the two cognates. In fact was a high pressure for the language spoken in Bessarabia to change, and the Russian influence led to massive changes in the vocabulary, and consequently the similarity dropped significantly. However, the newspaper language preserved much of its identity.

1832-1856	75%	1920-1950	87%
1856-1877	68%	1950-1980	NA
1877-1912	68%	1980-1990	NA
1912-1920	86%	1990-2015	95%

Table 3. Similarity as classifier confidence

3 Conclusions and Further Research

This research presents a diachronic survey conducted to compare journalistic language changes in the Romanian language in terms of time evolution across four regions, Bessarabia, Moldavia, Wallachia and Transylvania. The results highlight major similarities and interesting differences in these collections of publications.

We investigated the problem of diachronic similarity between the mass-media, newspaper, between cognate languages. In particular we focused on the relation between Romanian (including the historical regions: Moldova, Transylvania and Wallachia) and Bessarabian which, started with a high level of similarity and they are again to a very high level of similarity. The method we described is based on statistical analysis of words distributions over epochs reflected on the Romanian printing press and a statistical classifier, SVM, for each epoch. The methodology is language independent and offers an objective quantification of the similarity degree between old Romanian variants.

As further work we plan to expand the methodology farther by including (i) more data, including from period 1945-1990, when in Bessarabia the Latin alphabet was outlawed and (ii) implementing a deeper language analysis using and other statistical classifier as LSTM (Long Short Term Memory) in order to choose the best classifier in diachronic studies. We would like to investigate the semantic similarity between cognates by employing a deep learning approach as well.

Acknowledgments

I would like to thank Dr. Octavian Popescu for his constant guidance, endless suggestions and encouragement and full support to finish this work.

References

[Biber, D., 1987]. D. Biber. *A textual comparison of British and American Writing*. American Speech, (62), pages 99-119, 1987.

[Carstensen, K.-U et al., 2009]. K.-U. Carstensen, C. Ebert, S. Jekat, H. Langer, and R. Klabunde (eds.). *Computerlinguistikund Sprachtechnologie: Eine Einführung*. Spektrum Akademischer Verlag, 2009.

[Cole, R. et al., 1998]. R. Cole, J. Mariani, H. Uszkoreit, G. V. Battista Varile, A. Zaenen, A. Zampolli, V. Zue (eds.) *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1998.

[Cristea, D. and Butnariu C., 2004]. D. Cristea, and C. Butnariu. *Hierarchical XML representation for heavily annotated corpora*. In: Proceedings of the LREC 2004 Workshop on XML-Based Richly Annotated Corpora, Lisbon, Portugal, 2004.

[Davies, M., 2013]. M. Davies. *Recent shifts with three non-finite verbal complements in English: Data from the 100-million-word Time corpus (1920s-2000s)*. In: Aarts, Close, Leech and Wallis (eds.) The verb phrase in English: Investigating recent linguistic change with corpora, Cambridge: Cambridge University Press. pages 46-67, 2013.

[Delmestri, A. and Cristianini, N., 2010]. A. Delmestri and N. Cristianini. *String Similarity Measures and PAM-like Matrices for Cognate Identification*. Bucharest Working Papers in Linguistics, 12(2), pages 71-82, 2010.

[Densusianu, O., 1902]. O. Densusianu. *Filologia Romanică în universitatea noastră*, Bucuresci, J. V. Socecu Editeur, page 23, 1902.

[Diaconescu, P., 1974]. P. Diaconescu. *Elemente de istorie a limbii române literare moderne*. Partea I. Probleme de

normare a limbii române literare moderne (1830–1880), Bucureşti, 1974.

[Eggins, S., Martin, J.R., 1997]. S. Eggins, S., J. R. Martin. *Genres and Register of Discourse*. In: Dijk, T.A.v. (ed.) Discourse as Structure and Process (Discourse Studies – A Multidisciplinary Introduction), Vol. 1, pages 231–232. Sage Publications, London, UK, 1997.

[Gîfu, D. et al., 2016]. D. Gîfu, M. Dascălu, Ș. Trăușan-Matu, and L. Allen.*Time Evolution of Writing Styles in Romanian Language* at the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016, 3-9 Apr. 2016, Konya, Turkey.

[Gîfu, D. and Simionescu, R., 2016]. D. Gîfu, and R. Simionescu. *Tracing Language Variation for Romanian* at the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016, 3-9 Apr. 2016, Konya, Turkey.

[Gîfu, D., 2015]. D. Gîfu. *Contrastive Diachronic Study on Romanian Language*. In: Proceedings FOI-2015, S. Cojocaru, C.Gaindric (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, pages 296-310, 2015.

[Gooskens, C., 2006]. C. Gooskens. *Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility*. In: Van de Weijer, J. & Los, B. (eds.). Linguistics in the Netherlands, 23, Amsterdam: John Benjamins, pages 101-113, 2006.

[Gooskens, C. et al., 2008]. C. Gooskens, K. Beijering & W. Heeringa. *Phonetic and lexical predictors of intelligibility*. International Journal of Humanities and Arts Computing 2 (1-2), pages 63-81, 2008.

[Jurafsky, D. and Martin, J. H., 2009]. D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2009.

[Leech, G. at al., 2009]. G. Leech, M. Hundt, C. Mair, and N. Smith. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press, 2009.

[Manning, C. D. and Schütze, H., 1999]. C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[Popescu, O. and Strapparava, C., 2013]. O. Popescu and C. Strapparava. *Behind the Times: Detecting Epoch Changes using Large Corpora*. In International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013, pages 347-355.

[Popescu, O. and Strapparava, C., 2014]. O. Popescu and C. Strapparava. *Time corpora: Epochs, opinions and changes*. Knowledge-Based Systems, 2014.

[Simionescu, R., 2011]. R. Simionescu. *UAIC Romanian Part of Speech Tagger*, resource on nlptools.info.uaic.ro, "Alexandru Ioan Cuza" University of Iași, 2011.

[Trandabăț D., et al., 2012]. D. Trandabăț, E. Irimia, V. Barbu Mititelu, D. Cristea, D. Tufiș. *The Romanian Language in the Digital Age*. In: White Paper Series, Georg Rehm and Hans Uszkoreit (eds.), Berlin, Springer, 2012.

[Tufiș, D., Filip, F. Gh., 2002]. D. Tufiș, D., F. Gh. Filip (eds.). *Limba română în Societatea informațională – Societatea Cunoașterii*, Ed. Expert, București, 2002.