Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems (NESUS PhD 2016)
Timisoara, Romania

Jesus Carretero, Javier Garcia Blas
Dana Petcu
(Editors)

February 8-11, 2016

The Analysis of Diachronic Variation in Romanian Print Press

Daniela Gîfu

Alexandru Ioan Cuza University, Faculty of Computer Science, 16, General Berthelot St., 700483, Iaşi
daniela.gifu@info.uaic.ro

Abstract

The paper describes a study based on diachronic exploration of Romanian texts in order to implement a technology for detecting automatically the morpho-lexical from 1840 to nowadays. The chosen timings put in evidence the language changes, describing, also, the phenomena related to the evolution of the Romanian language, especially, in print press. We define a complex methodology for recovering of old Romanian texts in two different spaces: Romania (until 1918, representing 3 countries, Moldova, Wallachia and Transylvania) and Republic of Moldavia, the last being a territory lost of Romania after the historic events. The aim of this survey it to analyse the morphology and lexical-semantics of Romanian language, based on important corpus starting with the middle of the 19th century until today, in order to compare them, emphasizing the language differences and similarities. This work could be of interest to lexicographers and computational linguistics specialists, who want to clarify the linguistic identity.

Keywords diachronic study, lexicon, morphosyntax, print press, WEKA.

I. Motivation

This research is anchored in diachrony (over the centuries, Romanian language has crystallized some structures which continue to be preserved as we show later) at the expense of synchrony, since today, despite language innovations (Coşeriu, 1997) appeared, things seems to be more stable (Ciompec, 1985). It is about how can we investigate the linguistic deviations that affect the multilingual Republic of Moldova in parallel with the Romanian language, using natural language processing (NLP) methodology for tracking diachronic changes from the middle of the 19th century?

II. Related work

Up to the 16th century almost all scientific writing in Europe was conducted in Latin. The construction and annotation of historical corpora is challenging in many ways (Lüdeling et al. 2005; Chiarcos et al., 2008; Claridge, 2008; Rissanen, 2008; Kytö, 2011; Kytö and Pahta, 2012, among many others).

In general, the creation of a parallel corpus of diachronic language is constituted by biblical texts, because the Bible is one of the earliest sizable coherent texts documented for many languages (especially European). The reason is obvious, the digital text is freely available in an unparalleled variety of languages and it has been repeatedly updated in different periods of time (Resnik et al., 1999) becoming very useful for comparative and diachronic studies. For instance, for older Germanic languages (Sukhareva and Chiarcos, 2014).

The diachronically and synchronically comparative studies of the Romance languages expose the presence of many similarities, especially in diachronic studies (Densuianu, 1902). Latin was the starting point, but issues about substratum, superstratum and adstratum which contributed to differentiate language were not set aside.

Contributions assigned to this section are closely related to the previous ones, as many of the ideas in Romance linguistics are also found in diachronic or diatopic study of the Romanian language. Linguists are known to call for language facts from the Romanesque

in order to explain some form and vice versa. We should mention contributions of Al. Rosetti (Rosetti, 1968; Rosetti et al., 1971), Iorgu Iordan (Iordan, 1975), Al. Graur (Graur, 1968), Valeria Guţu-Romalo (Guţu Romalo, 1972; 2005), Florica Dimitrescu (Dimitrescu, 1978, 1982), Marius Sala (Sala, 1998), Victor Iancu (Iancu, 2000), Narcisa Forăscu (Forăscu, 2001), Angela Bidu-Vrănceanu (Bidu-Vrănceanu, 1986), Theodor Hristea, (Hristea, 1984) followed by those of Adriana Soichiţoiu-Ichim (Stoichiţoiu-Ichim, 2001), Rodica Zafiu (Zafiu, 2001), Grigore Brâncuş (Brâncuş, 2004) or Adrian Chricu (Chircu, 2012).

Reading the studies published by our predecessors helped us to better perceive the differences occurring in the Romanian language, in the diachronicy and diatopic. Taking over the way how to interpret the language facts from them, our system is developed based on morphological and syntactical analysis of the words found in analyzed ancient texts as highlighted by the methodology proposed in this paper.

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; Tufiş & Filip, 2002; Cristea & Butnariu, 2004; Trandabă
ct et al., 2012, Gîfu, 2015). The development and use of software for natural language processing (NLP) highlight the defining aspects of the text (morphological and syntactic analysis, semantic analysis and, more recently, pragmatic analysis).

The similarities between languages are interesting for historical and comparative linguistics, as well as for machine translation and language acquisition as well. Scannell (2006) and Hajič et al. (2000) argue for the possibility of obtaining a better quality in translation using simple methods for very closely related languages. Koppel and Ordan (2011) studied the impact of the distance between languages on the translation product and conclude that it is directly correlated with the ability to distinguish translations from a given source language from non-translated text. It has been established that some genetically related languages have a high degree of similarity to each other, and its speakers are able to communicate without prior instructions (Gooskens, 2006; Gooskens et al., 2008).

The approach for the study of the evolution of Romanian language is focusing only on the orthographic similarity. The basis for this approach consists of the idea that phonetic alterations have an orthographic correspondent, thus an alphabetic character correspondences (Delmestri and Cristianini, 2010).

Different approaches have been used in previous case studies in order to assess the orthographic distance similarity between related words. Their accuracy has been investigated and compared (Frunza et al., 2005; Rama and Borin, 2014), but a clear conclusion could not be drawn with respect to which method is the most appropriate for a given task. Metrics will be used to determine the orthographic similarity between related words. For the moment, we have the syllabic similarities of the Romanian language in different geographic areas and periods of time, starting by the Ciobanu and Dinu works (Ciobanu and Dinu, 2014). They used orthographic metrics like: the edit distance, the longest common subsequence ratio, and the rank distance.

III. Thesis idea

This survey describes the work methodology, starting with two collections of publications (Romanian and Moldavian), written at the middle of the 19th century, in order to compare them, emphasizing the language differences. In this sense, a modular structure is presented, including text processing, extracting quotes, WEKA statistics, and language similarity computation. As an illustration of the possible synergies between diachronic textual resources and linguistic research, a diachronic architecture is described using statistical machine learning techniques to infer probabilistic context-sensitive rules for the automatic delimiting in time and space of unknown words.

This amount of parallel data is of crucial interest to philologists and comparative linguists. Out of this context, it is also important for aligned journalistic corpora with the most important Romanian language resources as DEX-online and eDTLR, the last being developed by the Romanian Academy and âĂIJAlexandru Ioan CuzaâĂİ University of Iaşi.

IV. Authors and Affiliations

Formatting the authors' names and their affiliation depends on the number of authors and the number of different affiliations. Both names and affiliations spread over both columns.

V. Conclusion and future work

Language was not and is not static but the feature that characterizes language is the dynamism, whether it focuses on internal processes of word formation or loanwords. We were able to successfully create a search system for unknown words, acting especially on old text fields, these facts representing a premiere for Romanian language. For elaboration, symbolic method was used, combining efficiently rules created manually and a carefully organized external collection of files. It has been used two instruments of the Faculty of Computer UAIC, thus proving their usefulness: morphological and syntactic Tagger (WebPosRo) and Graphical Grammar Studio, and also improving existing findings. This resource can be useful in other projects on the same topic, where you only need to import.

By collecting all the information from an important resource we generate a large corpus that can be easily used in this application, but also this may be a way to extend the variation of programs that will use it. In this case, all this work of collecting content in order to get a large database will influence the final output of the main application.

Using the Naïve Bayse classifier available in WEKA, we managed to implement a mechanism which can find the words region and the period of time with 91% of correctly classified instances.

In the future we want to apply a few metrics in order to determine the orthographic similarity between related texts from the same period of time, but different areas. Moreover, we plan to extend this analysis for other kind of texts (literature, for instance), and to combine the orthographic approach with semantic evidence for a wider perspective on Romanian language similarity.

Acknowledgment

I would like to thank NESUS for supporting this article.

References

[1] Bidu-Vrănceanu, A. Structura vocabularului limbii române contemporane, Bucureşti, 1986.

[2] Carstensen, K-U., Ebert, C., Ebert, C., Jekat, S., Langer, H. and Klabunde, R. (eds.). ComputerlinguistikundSprachtechnologie: EineEinführung. Spektrum Akademischer Verlag, 2009.

[3] Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. Traitment automatique des langues, 49, 2008, pp. 271-293.

[4] Chircu, A. Influenţa slavă asupra limbii române pe baza ALRM I. Terminologia corpului omenesc. Harta 1 (Corp), în Katalin Balazs, Ioan Herbil (eds.), Lucrările simpozionului internaţional âĂđDialogul slaviştilor la începutul secolului al XXI-leaâĂİ (Cluj-Napoca, 8-9 aprilie 2011), Cluj-Napoca, Casa Cărţii de ştiinţă, 2012, pp. 92-98.

[5] Ciobanu, A. and Dinu, L. An Etymological Approach to Cross-Language Orthographic Similarity. Appilcation on Romanian in Proceedings of EMNLP-2014, Oct. 25-29, 2014, Doha, Quatar, pp. 1047-1058

[6] Ciompec, G. Morfosintaxa adverbului românesc. Sincronie şi diacronie, Bucureşti, Editura Ştiinţifică şi Enciclopedică, 1985, p. 283.

[7] Claridge, C. Historical Corpora. In A. Lüdeling, & M. Kytö (Eds.), Corpus Linguistics. An International Handbook, Volume 1. Berlin: De Gruyter, 2008, pp. 242âĂŞ259.

[8] Cole, R., Mariani, J., Uszkoreit, H., Battista V., Giovanni, Zaenen, Annie and Zampolli, Antonio (eds.). Survey of the State of the Art in Human Language Technology. Cambridge University Press, 1998.

[9] Coşeriu, E. Sincronie, diacronie şi istorie. Problema schimbării lingvistice, versiune în limba română de Nicolae Saramandu, BucureŞti, Editura Enciclopedică, 1997.

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

[11] Delmestri, A. and Cristianini, N. String Similarity Measures and PAM-like Matrices for Cognate Identification. Bucharest Working Papers in Linguistics, 12(2), 2010, pp. 71âĂŞ82.

[12] Densusianu, O. Filologia Romanică în universitatea noastră, Bucureşti, J. V. Socecu Editeur, 1902, p. 23.

[13] Dimitrescu, Florica (coord.). Istoria limbii române, BucureŞti, Editura Didactică Şi Pedagogică, 1978.

[14] Dimitrescu, Florica. Dicţionar de cuvinte recente, Bucureşti, Editura Albatros, 1982.

[15] Forăscu, N. Dificultăţi gramaticale ale limbii române, Ed. Univ., Bucureşti, 2001.

[16] Frunza, O., Inkpen, D., and Nadeau, D. A text processing tool for the Romanian language. Proceedings of the EuroLAN 2005 Workshop on Cross-Language Knowledge Induction, 2005.

[17] Gîfu, D. 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, 2015, pp. 296-310.

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

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

[20] Graur, Al. TendinÅčele actuale ale limbii române, Ed. Ştiinţifică, BucureŞti, 1968.

[21] Guţu Romalo, V. Corectitudine Şi greŞeală. (Limba română de azi), Bucureşti, 1972.

[22] Guţu-Romalo, V. Aspecte ale evoluţiei limbii române, col. "Repere", Bucureşti, Editura Humanitas Educaţional, 2005.

[23] Hajič, J., Hric, J., and Kuboň, V. Machine translation of very close languages. In Proceedings of the 6th Applied Natural Language Processing Conference, pages 7âĂŞ12. Association for Computational Linguistics, 2000.

[24] Hristea, Th. Sinteze de limba română, Editura Albatros, 1984.

[25] Iancu, V. Istoria limbii române, col. "Argumente", BucureŞti, Editura Fundaţiei Culturale Române, 2000.

[26] Iordan, I. Stilistica limbii române, Ed. Ştiinţifică, BucureŞti, 1975.

[27] Kytö, M. Corpora and historical linguistics. Revista Brasileira de Linguistica Aplicada, 11(2), 2011, pp. 417-457.

[28] Kytö, M., & Pahta, P. Evidence from historical corpora up to the twentieth century. In T. Nevalainen, & E. C. Traugott (Eds.), The Oxford Handbook of the History of English. Oxford o.a.: Oxford University Press, 2012, pp. 123-133.

[29] Lüdeling, A., Poschenrieder, T., Faulstich, L. C. et al. DeutschDiachronDigital - Ein diachrones Korpus des Deutschen. Jahrbuch für Computerphilologie 2004, 2005, pp. 119-136.

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

[31] Rama, T and Borin, L. Comparative Evaluation of String Similarity Measures for Automatic Language Classification. In George K. Mikros and Jan Macutek, editors, Sequences in Language and Text. De Gruyter Mouton, 2014.

[32] Resnik, P., Broman Olsen, M. and Diab, M. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. Computers and the Humanities 33, 1999, pp. 129-153.

[33] Rissanen, M. Corpus linguistics and historical linguistics. In: Corpus Linguistics: an International Handbook. Vol. 1, ed. by Anke Lüdeling and Merja Kytö. Berlin and New York: Walter de Gruyter. 2008, pp. 53-68.

[34] Rosetti, Al. Istoria limbii române, de la origini până în secolul al XVII-lea, cu 6 hărţi afară din text, BucureŞti, Editura pentru literatură, 1968.

[35] Rosetti, Al., Cazacu, B., Onu, L. Istoria limbii române literare, Bucureşti, Editura Minerva, 1971.

[36] Sala, M. De la latină la română, col. "Limba română", nr. 1, Bucureşti, Editura Univers Enciclopedic & Academia Română, 1998.

[37] Scannel, K. Statistical models for text normalization and machine translation. In Proceedings of the First Celtic Language Technology Workshop, pages 33âĂŞ40, Dublin, Ireland, August 23 2014.

[38] Stoichitoiu-Ichim, A. Vocabularul limbii romane actuale. Dinamica, influente, creativitate, Bucureşti, Editura All, 2001.

[39] Sukhareva, M. And Chiarcos, C. Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic in Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, Dublin, Ireland, August 23, 2014, pp. 11âĂŞ20.

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

[41] Tufiş, D., Filip, F. Gh. (coord.). Limba română în Societatea informaţională âĂŞ Societatea Cunoaşterii, Ed. Expert, Bucureşti, 2002.

[42] Zafiu, R. Diversitate stilistică în româna actuală, BucureŞti, 2001.