Time Evolution of Writing Styles in Romanian Language

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

Mihai Dascalu, Stefan Trausan-Matu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
{mihai.dascalu, stefan.trausan}@cs.pub.ro

Laura K. Allen
Institute for the Science of Teaching & Learning, Arizona State University
Tempe, USA
LauraKAllen@asu.edu

***Abstract*—This paper presents a diachronic analysis centered on the exploration of differences between the writing styles of journalistic texts in Romanian language. This analysis is focused on the time evolution of this language across two adjacent regions, Bessarabia and Romania in two major periods that were marked by important historical differences. Our aim is to examine these language differences based on corpora of historical and contemporary texts. To this end, we employ the ReaderBench framework to calculate a number of textual complexity indices that can be reliably used to characterize writing style. These analyses are conducted on two independent corpora for each of the two language styles, covering the following time periods: 1941-1991, when Bessarabia was separated from Romania and became a state in the Soviet Union (and there were few connections and language influences with Romania), and after July 1991, when Bessarabia became an independent state, Republic of Moldavia (and many language interactions with Romania occurred). The results of our analyses highlight the lexical and cohesive textual complexity indices that best reflect the differences in writing style, ranging from sentence and paragraph structure to word entropy and cohesion, measured in terms of Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).**

Keywords—writing style; language similarity; textual complexity; comparable corpora; time periods and geographic regions

I. INTRODUCTION

Language can be considered an archive similar to a "treasury (trésor)" that consists of grammatical acts that are developed through communication among individuals from similar communities [1]. The task of quantifying similarities among languages and dialects has become a major area of research in the field of Natural Language Processing (NLP) because typical approaches often requires highly skilled annotators and can be extremely time consuming. Prior research has proposed methods for creating sets of comparable corpora [1-6] that contain similar texts across multiple languages, which can then be used to assess linguistic differences among the languages. This comparable corpora approach has become a popular alternative to parallel corpora in diverse NLP tasks because larger volumes of data can be gathered faster, leading it to be less expensive and more productive. In this paper, we introduce a method for automatically comparing writing styles based on comparable corpora and we validate the method using journalistic texts written in a language spoken in two regions with the same language root: Romanian.

In linguistics, *diachrony* focuses on changes in word meanings over time, typically within the context of historical events. These lexical changes represent a topic of great interest for linguists, historians, and NLP researchers who use diachronic analyses to investigate the complexity of language, particularly as it is influenced by different historical factors. The analysis of *lexical fields* represents the continuation of Saussure's goal to provide consistency to the concept of linguistic value, particularly based on associative relationships between meanings (an objective since 1930) [7]. According to Saussure, diachronic approaches to linguistics focus on the *evolution* and *development* of language throughout history, whereas synchronic approaches do not take history into account [8]. Our research is anchored in this diachronic approach. In particular, we investigate the Romanian language, which is comprised of certain grammatical structures that have been preserved over centuries.

In order to explore diachrony in the Romanian language, we propose a method that relies on the comparison and contrast of writing styles among text according to a number of different indices: *text features* (e.g., length, structure or use of punctuation) [9], *textual formality* (e.g., vocabulary, slang, phrasal verbs, use of idiomatic language, personal pronouns or expression of attitude) [10], and *textual styles* (e.g., simple/complex sentences, stylistic markers, cohesion, reported speech or elliptical formulations) [11]. In the Romanian language, analyses of writing styles are not singular as they became constituent parts of the current trends in the interpretation of language facts [12-16]. In terms of vocabulary, language remains an inexhaustible source, even though many lexical units are ephemeral creations.

Even from the beginning we must emphasize that this study represents the first in-depth, large-scale experiment for which the Romanian textual complexity model was implemented and validated within our *ReaderBench* framework [17-19]. Although this framework previously covered English [17, 20] and French [21] languages, the development of the semantic models (LSA and LDA), of the Romanian Natural Language Processing pipeline and of the specific linguistic indices are also novel elements presented in detail in this paper. To our knowledge, the described approach highlights the most advanced, multi-hierarchical, automated discourse analysis

model of writing style existent for Romanian language covering more than 100 textual complexity indices.

The automated evaluation of textual complexity and writing styles represents a key focus among linguistics researchers and emphasizes the importance of technology to facilitate research on language. However, measuring textual complexity is a difficult task because automated indices must take into account the fact that perceptions of text difficulty can be influenced by numerous contextual variables such as prior knowledge, familiarity with the language, and personal motivation and interest. The ease with which a text can be comprehended is not only related to its linguistic structure, but also to the reader's education, cognitive capabilities and background experiences. Nonetheless, a number of systems for measuring text complexity have been developed [22], such as *Lexile* (MetaMetrics), *ATOS* (Renaissance Learning), Degrees of Reading Power: *DRP Analyzer* (Questar Assessment, Inc.), *REAP* (Carnegie Mellon University), *SourceRater* (Educational Testing Service) and *Coh-Metrix* (University of Memphis). Our framework, *ReaderBench* [17, 20], integrates the most common indices from these previous systems and places a strong emphasis on cohesion and semantics through the inclusion of additional indices, described in detail later in this paper [23]. Moreover, ReaderBench is the first framework to implement a multi-layered textual complexity assessment model [24] for the Romanian language.

II. Background and Current Study

The accuracy of linguistic similarity measures remains an open question in NLP. The literature points to a number of methods, each with their corresponding advantages and disadvantages. The most notable of these methods are: a) semantic distances [25] in lexicalized ontologies - WordNet [26] and Romanian version RoWN (*Romanian WordNet*) [27]; b) vector space models - *Latent Semantic Analysis* (LSA) [28] or word2vec [29]; c) topic models - *Latent Dirichlet Allocation* (LDA) [30]; d) machine translation evaluation - BLEU (*BiLingual Evaluation Understudy*) [31] or *Meteor* [32]; e) collocation extraction and associations between words - PMI (*Pointwise Mutual Information*) [33].

In this paper, we present a diachronic analysis centered on the exploration of differences between the writing styles of journalistic texts in the Romanian language from two adjacent regions: Bessarabia and Romania. Both dialects have a common origin, but are spoken in two different regions and have important historical differences. Our aim is to examine these language differences based on automated linguistic analyses of historical and contemporary texts. To this end, we employ the *ReaderBench* framework to calculate a number of textual complexity indices that can be reliably used to characterize writing style.

The analyses are conducted on two independent corpora for each of the two Romanian dialects, covering the following time periods: 1941-1991, when Bessarabia was separated from Romania and become a state in the Soviet Union (and there were few connections and language influences with Romania), and after July 1991, when Bessarabia became an independent state, Republic of Moldavia (and many language interactions with Romania occurred). Due to historical constraints, Romania can be considered the mother region, in contrast to the Bessarabian dialect. Although the language in Romania can be considered the baseline, we opted to use an analogy to two dialects emerging from the same original root language in order to facilitate a comparative view. Our specific research questions are below:

- Do journalistic texts produced in both Bessarabia and Romania during earlier (1941-1991) or later (1992-present) time periods differ in their *lexical complexity*?
- Are regional and time period differences detected in the *cohesion* and *semantics* of these texts?

In the remainder of this paper, we address these questions through a computational analyses of corpora written in the Romanian language. We will first describe the method for our current study, including a discussion of our assessment framework – *ReaderBench*. Next, we will offer an explanation and interpretation of the results of our diachronic analyses. Finally, we will conclude with a broad discussion of these results and describe directions for future work.

III. Method

In this section we present an overview of the two text collections analyzed in this study. These corpora belong to two time periods from Romania and the Bessarabia region (currently Republic of Moldova). Additionally, we describe the textual complexity indices that were calculated by *ReaderBench* to characterize the writing style of these texts.

A. Corpus selection

Our corpus was developed based on a newspaper collection and contains around 60,000 lexical tokens (see Table 1 for descriptive statistics). Importantly, this corpus represents a first iteration towards building a Romanian Gold corpus centered on diachronic meta-annotation.

Our analysis is constrained to two specific time periods (rather than earlier timeframes) for a number of reasons. First, after the second World War, the Cyrillic alphabet was enforced in Bessarabia by the USSR instead of the Latin script and there is no direct transliteration from one alphabet to another. Second, the LSA and LDA models integrated in *ReaderBench* (described in Section III.B) would have been inadequate due to important differences in word structure (old, obsolete forms not present in newer vocabularies), as well as completely different topics of interest and corresponding concepts. Third, access to texts from earlier timeframes is extremely limited and scarce. In order to keep the categories balanced, we have selected around 30 contemporary journal numbers from the four categories. Notably, we collected a higher number of Romanian texts from 1941-1991 because we wanted to make use of all the journal articles we had at our disposal. Although the majority of texts from Bessarabia were written using the Cyrillic alphabet during 1941-1991, our Latin selection is representative of the time period as the selected sources maintained their Romanian origins.

TABLE I. GENERAL CORPUS STATISTICS.

Region	Period	N documents	N words	Sources
Bessarabia	1941-1991	36	13,534	Basarabia, Curierul, Deşteptarea, Literatură şi artă
	1992-present	31	13,471	Contrafort, Jurnal, Jurnal de Chişinău, Literatură şi artă, Moldova suverană, Ziarul de gardă
Romania	1941-1991	63	25,873	Deşteptarea, România literară, Scânteia, Convorbiri literare, Moldova socialistă, Vatra românească
	1992-present	31	8,751	Dimineaţa copiilor, Evenimentul zilei, Gândul, Ziua, Ziua news, Jurnalul naţional
Total		137	61,629	

B. Textual complexity indices as markers of writing style

For the purposes of our analyses, the textual complexity indices calculated by ReaderBench were split into two primary categories:

- Lexical (e.g.,, average word length in characters, average number of unique content words per sentence, word entropy, average distance between lemma and word stems, and average distance between words and corresponding stems)
- Cohesion and Semantics (e.g., average paragraph-document cohesion – LSA, average paragraph-document cohesion – LDA, average intra-paragraph cohesion – LDA, and average transition cohesion – LDA)

1) Lexical indices

Early research on textual complexity was centered on the idea that computers can be used to automatically score student essays as effectively as expert human raters using only statistically and easily detectable attributes [34, 35]. The foundation of our model is derived from these metrics that have been used to automatically score essays. Additionally, it takes into consideration Slotnick's method [35, 36] of grouping proxies (computer approximations of interest) based on their intrinsic values. Therefore, *ReaderBench* includes a number of lexical indices from various categories, such as: average length of words, sentences and paragraphs in terms of characters, average number of content words per sentence and paragraph, average number of commas per sentence and paragraph, and the average distance between words, their corresponding lemmas or their stems. These indices are calculated based on content words, which are the lemmas of dictionary concepts that are not within the stop-words list.

In addition, measures of entropy [37] provide relevant insights into textual complexity at the character and word levels by measuring their diversity. This assumption of complexity relies on the following hypothesis: a more complex text contains more information, thus it requires more time and memory resources to process. Therefore, entropy measures of textual complexity reflect the diversity of characters and word stems found in a text. ReaderBench calculates the entropy of word stems, rather than actual words, because diversity at the syntactic level is better approximated using the root form of related concepts.

2) Cohesion and semantics

As highlighted by McNamara, et al. [38], textual complexity is strongly related to cohesion, and can have important effects on comprehension. In order to understand a text, the reader must develop a well-connected representation of the information they have read, which is often referred to as a situation model [39]. This connected representation is based on linking text fragments that occur throughout the text. Therefore, in *ReaderBench,* cohesion is reflected in our Cohesion Network Analysis approach [40] as the strength of inner-paragraph and inter-paragraph links. The structure of the cohesion graph influences readability as semantic similarities govern the understanding of a text.

Cohesion, from a computation perspective, relies on two semantic models – LSA (*Latent Semantic Analysis*) and LDA (*Latent Dirichlet Allocation*). LSA [28, 41] uses a training corpus to create a term-document matrix that contains a normalized number of occurrences for each word in a given document. The dimensionality of this matrix is reduced by applying Singular Value Decomposition, and words and documents are compared using a cosine distance between their vector representations in the projected semantic space. LDA [30, 42] is a generative probabilistic model based on topic distributions. Each topic is a Dirichlet distribution [43] over the vocabulary simplex (the space of all possible distributions of words from the training corpora) in which thematically related concepts have similar occurrence probabilities. Both models are based on the bag-of-words approach and reflect co-occurrence patterns emerging from the training corpora. In the current study, the LSA and LDA semantic models were trained on a Romanian corpus of more than 2 million content words covering a wide range of linguistic registers, such as journalistic, literature, science, and religion, with different social origins (e.g., suburban language or slang).

C. Statistical analyses

Statistical analyses were conducted to investigate differences in the writing styles of journalistic texts based on the region and time period in which they were produced. As mentioned in the previous section, our analyses focused solely on the lexical and cohesive properties of the texts.

We separately conducted statistical analyses on the two groups of linguistic indices (i.e., lexical and cohesive). First, the Shapiro-Wilk test of normality was conducted to assess normal distributions of the indices reported by *ReaderBench*. All variables that demonstrated non-normality were removed from the analysis. Multicollinearity of the variables was then assessed as pair-wise correlations ($r > .80$). In the case that indices demonstrated multicollinearity, the index that demonstrated the strongest effect in the model was retained for the final analysis (see Table 2 for final list of indices and their descriptive statistics). Finally, two multivariate analyses of

variance (MANOVAs) were conducted to examine whether the *lexical* and *cohesion and semantics* properties of the texts differed across region and time period.

IV. RESULTS AND DISCUSSIONS

This section describes the MANOVA analyses used to highlight the differences in writing styles of journalistic collections based on region (Romania and Besserabia) and time period.

TABLE II. GENERAL STATISTICS.

Index	Region	1941-1991 M (SD)	1992-present M (SD)
Lexical indices			
Average word length in characters	Bessarabia	3.95 (0.39)	4.38 (0.39)
	Romania	3.89 (0.42)	4.13 (0.40)
Average number of unique content words per sentence	Bessarabia	5.80 (2.40)	8.20 (1.89)
	Romania	6.95 (3.09)	7.41 (2.23)
Word entropy	Bessarabia	4.87 (0.30)	4.73 (0.46)
	Romania	4.89 (0.39)	4.65 (0.20)
Average distance between lemma and word stems (only content words)	Bessarabia	0.85 (0.16)	0.93 (0.13)
	Romania	0.85 (0.16)	0.9 (0.17)
Average distance between words and corresponding stems (only content words)	Bessarabia	1.32 (0.26)	1.44 (0.30)
	Romania	1.23 (0.25)	1.35 (0.25)
Cohesion and Semantics			
Average paragraph-document cohesion (LSA)	Bessarabia	0.59 (0.16)	0.61 (0.11)
	Romania	0.63 (0.14)	0.60 (0.11)
Average paragraph-document cohesion (LDA)	Bessarabia	0.60 (0.13)	0.61 (0.07)
	Romania	0.64 (0.11)	0.63 (0.10)
Average sentence-paragraph cohesion (LSA)	Bessarabia	0.65 (0.12)	0.72 (0.12)
	Romania	0.61 (0.14)	0.68 (0.12)
Average intra-paragraph cohesion (LDA)	Bessarabia	0.44 (0.10)	0.49 (0.08)
	Romania	0.41 (0.12)	0.49 (0.07)
Average transition cohesion (LDA)	Bessarabia	0.44 (0.14)	0.47 (0.10)
	Romania	0.43 (0.10)	0.45 (0.09)

A. Lexical analyses

Our first research question regarded whether journalistic texts produced in both Bessarabia and Romania during earlier (1941-1991) or later (1992-present) time periods differed in their complexity at the word-level. A MANOVA was conducted to examine the differences in the lexical indices across the two regions and time periods (see Table 2 for descriptive statistics). No two lexical indices correlated above r = .80; therefore, no indices were removed from the analysis.

This analysis revealed that there was a main effect of region [$F(5, 153) = 2.38$, $p < .05$] and time period [$F(5, 153) = 5.212$, $p < .001$], as well as a significant interaction between these two factors [$F(5, 153) = 2.41$, $p < .05$]. Texts from Bessarabia were characterized by higher lexical sophistication, both in terms of longer words, $F(1, 1.30) = 6.74$, $p = .01$, and marginally more elaborated word inflections (*average distance between words and corresponding stems*), $F(1, 0.29) = 3.83$, $p = .052$, compared to Romanian texts.

In terms of time period, later texts (1992-present) were more lexically sophisticated than earlier texts, comprising longer words [$F(1, 3.73) = 19.35$, $p < .001$], higher incidences of unique content words [$F(1, 81.12) = 8.59$, $p < .01$], and greater distance between lemma and word stems [$F(1, 0.11) = 4.01$, $p < .05$], as well as between words and corresponding stems [$F(1, 0.40) = 5.28$, $p < .05$]. Conversely, later texts contained lower word entropy compared to earlier texts, $F(1, 0.70) = 4.49$, $p < .05$.

Finally, the number of unique content words per sentence exhibited a significant interaction with period and region [$F(1, 1.30) = 6.74$, $p = .01$], indicating that the unique words in Romanian texts did not differ across time, whereas Bessarabian texts increased from the earlier to later time periods.

B. Cohesion and semantic analyses

A second MANOVA was conducted to examine differences in the cohesion and semantics indices across the two regions and time periods (see Table 2 for descriptive statistics). Two of the indices correlated above $r = .80$: average paragraph-document cohesion (LSA) and average paragraph-document cohesion (LDA); therefore, the index that demonstrated the smallest effect [average paragraph-document cohesion (LSA)] was removed from the analysis. The analysis revealed that there was a main effect of time period [$F(4, 130) = 4.79$, $p < .001$], but no main effect for region ($p = .20$), nor was there a significant interaction between the two factors ($p = .77$). Later texts (1992-present) exhibited greater sentence-paragraph cohesion [$F(1, 3.73) = 19.35$, $p < .001$] and intra-paragraph cohesion [$F(1, 0.12) = 12.16$, $p = .001$]. Thus, the results of this analysis suggest that semantic cohesion increased in texts over time, but did not differ according to region.

The MANOVA analyses provide evidence that the writing styles of journalistic texts from different regions and time periods were significantly different at the lexical and cohesion levels. However, differences were most strongly reflected in the lexical properties of the texts, rather than the cohesion. In order to better observe the differences in the evolution of the Romanian language, comparative charts are provided in Figure 1.

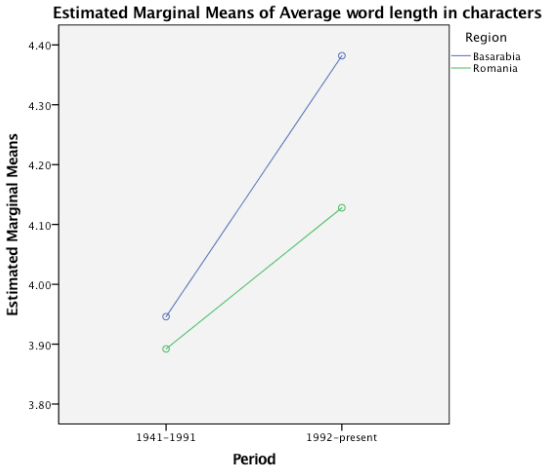
V. DISCUSSION OF RESULTS

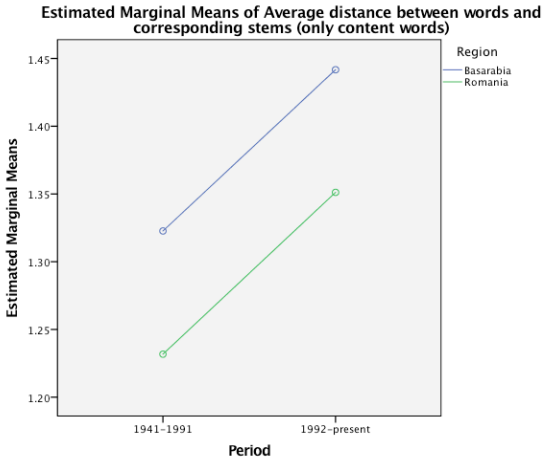
At the word level, a clear pattern can be observed, as texts from both regions increase in their lexical sophistication, with words increasing in their overall length and in the length of their suffixes and/or prefixes (see Figure 1.a). The enrichment of Romanian language with western terminology, especially English and French concepts was a natural process after the fall of communism in December 1989 and the borders were opened. As an example, before 1990 the word "*miliție*" from "*milicÿa*" (Russian) was used, whereas after 1992 a switch occurred towards "*poliție*", equivalent to "*police*" (English, French) or "*polizei"* (German). Another interesting and remarkable example is "*bolșevic*" from "*bol'ševik*" (Russian), which after 1992 became "*comunist*" derived from "*communiste*" (French) or "*communist*" (English). Another contributing element after the 1990s is the use of prefixes in both languages, especially after returning to the Latin script for the Bessarabian dialect. For example, "*paraștiințifice*" (En.

"*scientific fiction*") is a word composed of the prefix "*para*" and "*științifice*" (En. "*scientific*").

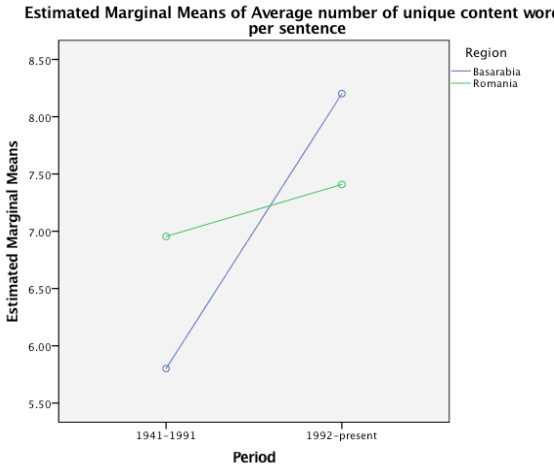
In terms of vocabulary, phrases tended to become longer, including more content words, but there was a lower overall diversity of concepts, as evidenced by decreases in word entropy (see Figure 1.b). Finally, local cohesion measured by LSA and LDA increased as texts written in the later time period were written with more self-contained, cohesive paragraphs (see Figure 1.c).

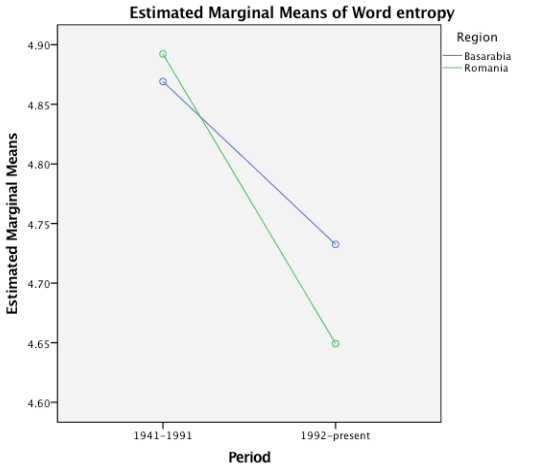
(a)



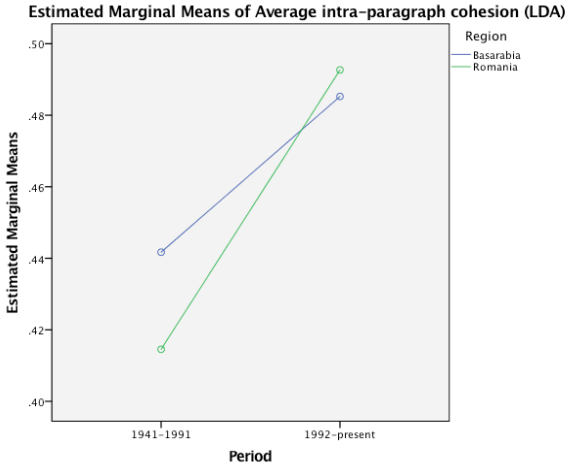


(b)





(c)



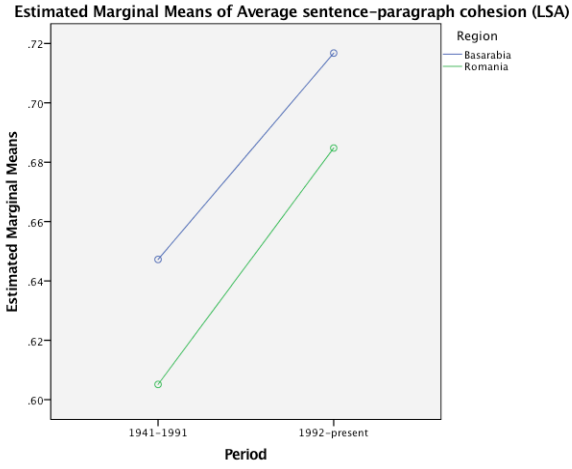


Fig. 1. Comparative views of the Romanian language time evolution in both regions.

In comparison to the Romanian language and coupled with historical events, the Bessarabian dialect had a tense history. After Bessarabia was occupied during World War II by the USSR, the Latin script was replaced (until 1991) by the Cyrillic alphabet and the Romanian language from that region was highly affected. The excessive use of the Russian language

resulted in the degradation of the Romanian language in this territory primarily in the context of spoken language. Some writings still used the Latin script (e.g., the ones considered in our study), but they were scarce.

In addition, it is worth mentioning that the Bessarabian vocabulary, particularly before 1991, is a combination of archaisms (i.e., Romanized words with Russian roots), most of them from the eastern region of Romania (Moldova), and words/sintagms derived from Russian words. As an example, the sentence "*M-ai ubidit, tovarăşe*!" is structured as follows: "*M-ai*" (En. "*me*") + "*ubidit*" (En. "*to see*", archaic form of Romanian verb "*a (se) vedea*" derived from the Russian root, "*уʙuдum*") + "*tovarăşe*", a word assimilated with the communist regime, which has been replaced with "*domnule*" (En. "*mister*")

As previously mentioned, a lexical enrichment due to the communication freedom and the Internet usage can be observed after the 1990s in both regions, with most concepts being influenced by English. For example, the sentence "*Nu mă pot focusa*!" is structured as follows: "*Nu*" (En. "*not*") + "*mă*" (En. "*me*") + "*pot*" (En. "*can*") + "*focusa*" (a Romanian new word with English root, "*to focus*"). Moreover, texts tended to become more complex with longer sentences usually with 3 to 5 clauses.

VI. Conclusions and Future Work

This research presents a comparative study of texts written in the Romanian language in terms of their time evolution across two regions, Bessarabia and Romania. The results reveal important and interesting differences in these texts. Moldavian texts were revealed to be higher in their lexical sophistication, with longer words and with marginally more elaborated inflections as compared to Romanian texts. Similarly, texts written after 1992 were more elaborated; they contained longer words and longer suffixes and prefixes. The number of unique content words in Bessarabian texts increased from the earlier to later time periods while Romanian texts did not differ across time.

The semantic cohesion increased in the texts over time, but did not differ according to region, with only one exception. For the moment, these parameters denote a clear pattern of increased elaboration over time in the texts from both regions. Importantly, these similarities between writing styles are not surprising given the historical facts that Republic of Moldavia was part of Romania from 1918 to 1941, and became an independent state after 1991. Of course, important differences were also detected, perhaps due to the influence of the Russian language reflected on the Romanian language in Bessarabia, starting from the middle of the 19th century.

Our aim is to further extend the collection of documents with transliterated texts from the Cyrillic alphabet in order to increase the size and representativity of our corpus. We were constrained by the scarcity of available documents and our goal was to have a balance between periods and regions within this pilot study.

In addition, further research should be conducted to more explicitly understand the causes of these results and to identify additional text sources in order to increase the power of the statistical analysis. Another potential distinction that should be investigated is between texts written in the three different regions of Romania (Wallachia, Moldavia, Transylvania) versus Bessarabia.

Acknowledgements

This work has been partially funded by the 2008-212578 LTfLL FP7 project, as well as by the EC H2020 project RAGE (Realising and Applied Gaming Eco-System); http://www.rageproject.eu/; Grant agreement No 644187.

References

[1] F. de Saussure, *Cours de linguistique générale*. Paris: Payot, 1999.

[2] L. Bo, E. Gaussier, E. Morin, and A. Hazem, "Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue," in *Conf´erence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France, 2011, pp. 211–222.

[3] E. Morin and B. Daille, "Comparabilité de corpus et fouille terminologique multilingue," *Traitement Automatique des Langues,* vol. 47, pp. 113–136, 2006.

[4] D. Gîfu, "Contrastive Diachronic Study on Romanian Language," in *FOI-2015*, 2015, pp. 296–310.

[5] J. Preiss, "Identifying Comparable Corpora using LDA," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, 2012, pp. 558–562.

[6] K. Aijmer, B. Altenberg, and M. Johansson, *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994* vol. 88: Lund studies in English, 1996.

[7] F. de Saussure, *Recueil des publications scientifiques: Ferdinand de Saussure*. Genève: Société anonyme des éditions Sonor, 1922.

[8] E. Coşeriu, *Sincronie, diacronie şi istorie. Problema schimbării lingvistice*. Bucureşti, Romania: Editura Enciclopedică, 1997.

[9] National Governors Association Center for Best Practices & Council of Chief State School Officers, "Common Core State Standards," ed. Washington D.C.: Authors, 2010.

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

[11] D. Biber, "A textual comparison of British and American Writing," *American Speech,* pp. 99–119, 1987.

[12] A. Rosetti, B. Cazacu, and L. Onu, *Istoria limbii române literare*. Bucureşti, Romania: Editura Minerva, 1971.

[13] I. Iordan, *Stilistica limbii române*. Bucureşti, Romania: Editura Ştiinţifică, 1975.

[14] F. Dimitrescu, Ed., *Istoria limbii române*. Bucureşti, Romania: Editura Didactică şi Pedagogică, 1978, p.^pp. Pages.

[15] M. Sala, *De la latină la română* vol. 1 of Limba română. Bucureşti, Romania: Editura Univers Enciclopedic & Academia Română, 1998.

[16] V. Guţu-Romalo, *Aspecte ale evoluţiei limbii române* vol. Repere. Bucureşti, Romania: Editura Humanitas Educaţional, 2005.

[17] M. Dascalu, P. Dessus, M. Bianco, S. Trausan-Matu, and A. Nardy, "Mining texts, learner productions and strategies with ReaderBench," in *Educational Data Mining: Applications and Trends*, A. Peña-Ayala, Ed., ed Cham, Switzerland: Springer, 2014, pp. 345–377.

[18] M. Dascalu, L. L. Stavarache, P. Dessus, S. Trausan-Matu, D. S. McNamara, and M. Bianco, "ReaderBench: The Learning Companion," in *17th Int. Conf. on Artificial Intelligence in Education (AIED 2015)*, Madrid, Spain, 2015, pp. 915–916.

[19] M. Dascalu, L. L. Stavarache, S. Trausan-Matu, P. Dessus, M. Bianco, and D. S. McNamara, "ReaderBench: An Integrated Tool Supporting both Individual and Collaborative Learning," in *5th Int. Learning*

Analytics & Knowledge Conf. (LAK'15), Poughkeepsie, NY, 2015, pp. 436-437.

[20] M. Dascalu, *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* vol. 534. Cham, Switzerland: Springer, 2014.

[21] M. Dascalu, L. L. Stavarache, S. Trausan-Matu, P. Dessus, and M. Bianco, "Reflecting Comprehension through French Textual Complexity Factors," in *26th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2014)*, Limassol, Cyprus, 2014, pp. 615–619.

[22] J. Nelson, C. Perfetti, D. Liben, and M. Liben, "Measures of text difficulty: Testing their predictive value for grade levels and student performance," Council of Chief State School Officers, Washington, DC2012.

[23] M. Dascalu and D. Gifu, "Evaluating the Complexity of Online Romanian Press," in *11th Int. Conf. "Linguistic Resources and Tools for Processing the Romanian Language"*, Iasi, Romania, 2015, pp. 149–162.

[24] M. Dascalu, L. L. Stavarache, P. Dessus, S. Trausan-Matu, D. S. McNamara, and M. Bianco, "Predicting Comprehension from Students' Summaries," in *17th Int. Conf. on Artificial Intelligence in Education (AIED 2015)*, Madrid, Spain, 2015, pp. 95–104.

[25] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics,* vol. 32, pp. 13–47, 2006.

[26] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998, p.^pp. Pages.

[27] D. Tufiş, V. Barbu Mititelu, L. Bozianu, and C. Mihăilă, "Romanian WordNet: New Developments and Applications," in *3rd Global Wordnet Conference 2006 (GWC2006)* Jeju Island, Korea, 2006, pp. 337–344.

[28] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge," *Psychological Review,* vol. 104, pp. 211–240, 1997.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representation in Vector Space," in *Workshop at ICLR*, Scottsdale, AZ, 2013.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993–1022, 2003.

[31] K. Papieni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 311–318.

[32] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *EACL 2014 Workshop on Statistical Machine Translation*, Gothenburg, Sweden, 2014.

[33] K. W. Church and P. Hanks, "Word association norms, mutual information and lexicography," *Computational Linguistics,* vol. 16, pp. 22–29, 1990.

[34] E. Page, "The imminence of grading essays by computer," *Phi Delta Kappan,* vol. 47, pp. 238–243, 1966.

[35] W. Wresch, "The imminence of grading essays by computer—25 years later," *Computers and Composition,* vol. 10, pp. 45–58, 1993.

[36] H. Slotnick, "Toward a theory of computer essay grading," *Journal of Educational Measurement,* vol. 9, pp. 253–263, 1972.

[37] C. E. Shannon, "Prediction and entropy of printed English," *The Bell System Technical Journal,* vol. 30, pp. 50–64, 1951.

[38] D. S. McNamara, A. C. Graesser, and M. M. Louwerse, "Sources of text difficulty: Across the ages and genres," in *Measuring up: Advances in how we assess reading ability*, J. P. Sabatini, E. Albro, and T. O'Reilly, Eds., ed Lanham, MD: R&L Education, 2012, pp. 89–116.

[39] T. A. van Dijk and W. Kintsch, *Strategies of discourse comprehension*. New York, NY: Academic Press, 1983.

[40] M. Dascalu, S. Trausan-Matu, D. S. McNamara, and P. Dessus, "ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism," *International Journal of Computer-Supported Collaborative Learning,* vol. 10, pp. 395–423, 2015.

[41] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, pp. 259–284, 1998.

[42] D. M. Blei and J. Lafferty, "Topic Models," in *Text Mining: Classification, Clustering, and Applications*, A. Srivastava and M. Sahami, Eds., ed London, UK: Chapman & Hall/CRC, 2009, pp. 71–93.

[43] S. Kotz, N. Balakrishnan, and N. L. Johnson, "Dirichlet and Inverted Dirichlet Distributions," in *Continuous Multivariate Distributions*. vol. 1: Models and Applications, ed New York, NY: Wiley, 2000, pp. 485–527.