“ALEXANDRU IOAN CUZA” UNIVERSITY OF IAȘI
FACULTY OF COMPUTER SCIENCE

LEXICAL SEMANTICS IN TEXT PROCESSING

Contrastive Diachronic Studies on Romanian Language

- PhD thesis -

Supervisors:
Prof.univ.dr. Dan Cristea
Acad. Dan Tufiș

PhD student:
Daniela Gîfu

Iași
July 2016

Contents

ACKNOWLEDGMENTS

I would like to thank my thesis supervisors, professor Dan Cristea and professor Dan Tufiș, the creators of the schools of computational linguistics in Iași and Bucharest, for attent guiding during the whole period of PhD studentship that has finally resulted in this thesis. Since 2009, they have been my closest scientific collaborators, teaching me the secrets of the fields of computational linguistics and natural language processing and inspiring my research. I consider that the scientific endeavor that I have invested in doing this PhD research is the most important step in my professional life.

During the 3 years of Doctoral School and before them, I had the opportunity to be guided by admirable professors from the Faculty of Computer Science and the Faculty of Letters of "Alexandru Ioan Cuza" University of Iași, Corina Forăscu, Adrian Iftene and Dan Stoica. I thank them for their so many comments and corrections referring my PhD repports and different chapters of this thesis, for encouragement and fresh ideas at critical times.

I also express my indebted appreciation to Octavian Popescu, Radu Simionescu and Mihai Dascălu for their guidance and suggestions regarding different models and methods I have used in this research.

I express my vivid gratitude to my colleagues from the Academy of Sciences from Chișinău, especially to Svetlana Cojocaru, and the "Alecu Russo" State University in Bălți, especially to Valentina Prițcan and Elena Harconița, who have helped me to develop the Bessarabian corpus for the period 1944-1989, when the Latin script was banned in the Republic of Moldavia. The beginnings of my work to develop the corpus are linked to the many travelings I have made in Chișinău and Bălți between 2014-2016.

I thank also my friends, Irina Macovei and Augusto Perez, for the time and effort they have invested in proofreading parts of this PhD thesis.

It is appropriate to thank my colleagues Alina, Anca, Andrei, Daniel, Diana, Gabriela, Elena, Ionuț, Isabelle, Lavinia, Lăcrămioara, Lenuța, Lucian, Maria, Mădălin, Mihaela, Mircea, Mirela, Paul, Paula, Verginica, which have been close to me the whole period of my PhD studies.

Finally, I thank to my mom, my dad and my husband, the family who believed in me and

supported me being for such a long time so far away from them.

I am grateful to good God, who gives me dayly the energy to go forward.

1. INTRODUCTION

In the context of information society, a retrospective linguistic investigation is a major project, being essential in the architecture of an automatic language processing system. In terms of methodology, the temporal criterion becomes essential in recognizing language phenomena in a given space.

I have chosen for this PhD thesis a collection of journalistic texts from the period 1829-2015, a significant period in the affirmation and development of the Romanian and Bessarabian journalistic style, in order to identify the epochs when the lexic and the morphology for the written language of these areas are more or less similar. The entire journalistic context is discussed in Chapter 6.

The PhD title - **Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language** - announces the crossing of a very complex audacious process, with many aspects still unclear and certainly new in the area of Romanian computational linguistics. It is the evidence of an (inter)national teamwork where a few specialists from Romania, Bessarabia and USA were involved, whose merits will be mentioned along the way. Actually, we talk about the newspaper language analysis during the mentioned period, which required an impressive effort, especially in building a large diachronic corpus we want to develop. It is extremely necessary for future linguistics and computational linguistics approaches, both concerned with the understanding of language in written and spoken forms.

We believe that this corpus is representative for such an analytic demarche, aiming to recognize the language deviations from the current norm in certain epochs that will be automatically identified with different statistical methods, described in Chapter 5. That is why the level of technology of a language is closely related to the complexity and availability of language resources. The Romanian language resources are described in Chapter 4.

The rich literature speaks for itself about the usefulness of technologies and computational services, some of the more relevant previous studies for this PhD research being described in Chapter 2. The development and utilization of software programs for natural

language processing (NLP) highlight the defining aspects of the text (morphosyntactic analysis, semantic analysis and, more recently, pragmatic analysis).

Since the second half of the 20th century, the informatization of the Romanian language has become a priority for programmers linguists who define and implement various formal models with the desire to develop a significant linguistic processing automatic system.

Because the language is an extremely complex phenomenon, the diachronic aspects are analyzed using statistical models based on a large corpus of texts (here, newspapers), from which the morpho-syntactic and semantic features are extracted. Lastly, the norms of the journalistic language are part of a selection process, their evolution fits of a natural process that we want to analyze it.

1.1 Aim and Interests

The performances offered by the computer in NLP are undeniable, but they are based on a long and deep theoretical framework that aims especially the inherent specificities of the old newspaper language. In other words, the current diachronic analysis necessitates a major theoretical background that serves for measuring language similarities between cognate languages based on a large collection of newspapers. As shown in Romanian literature, the evolution of the Romanian printing press coincides with the evolution of the Romanian modern literary language [Diaconescu, 1974]. Basically, this work demonstrates the wider range of application possibilities which may include different types of texts (analytical essays, short stories, editorials, novels, etc.).

Novelty of this research consists of a new approach for a diachronic contrastive analysis: a set of machine learning techniques in order to measure language similarity between cognate languages since the second decade of the 19th century.

Our research is based on the premise that a deep language analysis for machine learning is best applicable in diachrony.

This work investigates the linguistic crisis that affects the newspapers language from Bessarabia and Romania, using natural language processing (NLP) methodology for tracking language changes over time. In this case, the methodology relied on part-of-speech (POS) tagging, often with manual post-editing. Most previous work [Cole *et al.*, 1998; Manning & Schütze, 1999; Tufiș & Filip, 2002; Cristea & Butnariu, 2004; Delmonte, 2007; Leech *et al.*, 2009; Carstensen *et al.*, 2009; Jurafsky & Martin, 2009; Trandabăț *et al.*, 2012; Davies, 2013,

and many more] has focused mainly on the linguistic interpretation of the results. Their hypotheses were based on the ways language changes, without considering the cause of the statistically results.

The aim of this PhD research is to examine the differences between the writing styles of journalistic texts in Romanian language based on corpora of historical and contemporary texts. This analysis is centered on the time evolution of this language across four adjacent regions, Bessarabia, Moldavia, Transylvania and Wallachia.

We consider that this work offers a promising basis for future large-scale studies, reducing the amount of human effort required. The choice of statistical tests was evaluated and confirmed to be promising by some procedures, relying on the text mining techniques and models as: Naïve-Bayes, SVM (Support Vector Machine), LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis), MaxEnt (Maximum Entropy), LSTM (Long Short-Term Memory), etc.

1.2 PhD Thesis Outline

In this PhD thesis, we propose a diachronic model derived from text mining. In addition, as a personal target, the second goal of this work was to induce a balance in the internal structure of each chapter, as well as between its different sections. Thus, the theoretical device, on the one hand and the practical device, on the other hand, complete the diachronic picture represented by the journalistic context analyzed in Chapter 6, text mining techniques and models in Chapter 5, in conjunction with the language similarity analysis (see Chapter 7).

Furthermore, we opted for six theoretical chapters, each with at least 3 sections to increase the stability of the theoretical framework of this thesis. Even though each chapter has a well-defined role, we were equally interested in the links between chapters. These have been well established with the aim to highlight the interdependencies of the major directions in relation to the main goal of PhD research as follows: 2.1 Historical corpora are connected to 6.1.1 *A brief history of Romanian newspapers. Romanian corpus description* and 6.2.2 *A brief history of Bessarabian newspapers. Bessarabian corpus description*; 2.2 *Diachronic text classification* creates an echo in 5.2.1 *Naïve Bayes – a generative model*; from 2.3 *Languages similarities using machine learning* described in 5.2.2 *Maximum Entropy Modeling*, 5.3.1 *Support Vector Machines* (SVMs) 5.4.1 *LDA – a probabilistic topic model* and 5.4.2 *LSA – a vector space model* used in the empirical studies for automatic analyses.

Therefore, the entire theoretical contents of this PhD thesis are orientated to the understanding of the tools role, necessary for an exceptional analytic demarche. In addition, the Chapter 8 punctually records the final remarks about each information from this work.

Briefly, this PhD thesis is structured as follows:

Chapter 2 addresses language is studying by firstly considering the multiple approaches of diachronic corpora and language changes over time, their connections to language similarity, also grounding a computational view. In order to provide a clearer perspective of the theoretical aspects, we categorize the literature works depending on the diachronic level analysed.

Chapter 3 introduces basic ideas from linguistics and computational linguistics to state the importance of lexical semantics in the analysis of texts.

Chapter 4 describes an inventory of the most known lexical resources in NLP, which we used for the purpose of this PhD thesis (e.g. Word**N**et lexical ontology), and the contrastive diachronic studies on Romanian language, which determines the degree of similarity between cognate languages.

Chapter 5 includes some techniques and models within the subdomain Natural Language Processing (NLP) named „Text Mining", being a knowledge-intensive process. They are directly related to the PhD research goal, namely to follow the language variation over time betwen cognate languages.

Chapter 6 presents the journalistic context from Romania and Bessarabia, on which we built the corpus of this PhD thesis. This corpus has a chronological logic, which the literature records it as the main indicator in the analysis of the evolution of written press, underlining alsoits main particularities.

Chapter 7 includes a detailed description of analytical scenarios envisaged, emphasizing both the advantages and the inherent limitations that result from our statistical results. Our diachronic methodology could be useful both for scientific goals (exploring the nature of cognate languages) and for applicative goals (language similarity analysis using statistical methods).
Finally, the Chapter 8 depicts succinct major contributions of this PhD research and reflections about it, together with future work.

The 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. Finally, this PhD thesis offers the basis of linguistic theory to build a language model that captures the deviations from the current norm, seen as a linguistic convention [Șuteu, 1976]. It manifests itself in all language departments, such as: in phonetics (pronouncing and writing of the sounds), in grammar (e.g. keeping flexion), in lexic (using those words understood throughout the

language territory of a state).

2. STATE OF THE ART

This chapter addresses language studying by firstly considering the multiple approaches of diachronic corpora and language changes, their links to language similarities, also grounding a computational view to be discussed in Chapter 6. The rich literature tells its own story regarding the usefulness of technology and information services: data reduction procedures [Carstensen *et al.*, 2009], speech and language processing [Jurafsky & Martin, 2009], statistical techniques for NLP (Natural Language Processing) [Manning & Schütze, 1999], human language technology [Cole *et al.*, 1998]; spoken language technology [Tufiș & Filip, 2002]; discourse technologies [Cristea & Butnariu, 2004]; semantic interpretation [Delmonte, 2007]; diachronic text classification [Szymanski & Lynch, 2015]; diachronic text evaluation [Popescu & Strapparava, 2013], etc.

In the context of the information society, a diachronic retrospective investigation is a major project, essential in the architecture of an automatic processing language system that we describe in Chapter 7. Many previous works [Baayen, 2008, Leech *et al.*, 2009; Davies, 2013] have focused mainly on the linguistic interpretation of the statistically results. The hypotheses of their works concerned the ways language changes based on statistical methods.

In order to provide a clearer perspective of the theoretical aspects, we categorize the literature works depending on the chosen approach with reference to the following aspects: building diachronic corpora, annotating diachronic corpora, diachronic text classification, similarity language using various statistical techniques, identifying epochs in the evolution of a language using statistical methods, etc.

2.1 Historical corpora

The analysis of language changes over time implies a set of multifactorial causes, described in [Biber & Conrad, 2009], which requires a large diachronic corpus or historical corpus and a deeply annotation. However, developing historical corpora requires more time and resources.

In [Claridge, 2008:250] there are presented the advantages and disadvantages of the use of printed editions for compiling historical corpora, but we must remember the facility of many

language analysis types.

Currently, the diachronic corpora of an increasing number of languages are constructed and annotated with syntactically (e.g. words are tagged with their POS - part-of-speech) relevant information in order to address them and related issues [Kytö, 2011; Kytö & Pahta, 2012].

The construction and annotation of historical corpora is a challenge for NLP field, in many ways. Here are some representative historical corpora.

English historical linguistics has a long-standing tradition of corpus based work. The largest historical structured corpus for English is described in [Davies, 2015]. It is abbreviated with COHA (Corpus of Historical American English) and contains more than 400 million words of text from the 1810s-2009s, being balanced by genre decade by decade. COHA is tagged for POS, including for verbs and the particles of phrasal verbs. It includes a number of various registers: fiction, newspapers, non-fiction books, etc.

In addition, there are two more important corpora for English used to explore the diachronic phenomena. In fact, using historical corpora we can identify unknown correlations between language usage and different time periods. In [Michel *et al.*, 2011] is presented first version of the Google Books Ngram Corpus[1] used for different quantitative linguistic analysis. It contains millions of books in the last five centuries, giving us syntactically annotated ngrams. It is structured by words and phrases (i.e. ngrams) for eight languages (English, Spanish, French, German, Russian, Italian, Chinese, and Hebrew) and their frequency over time. Another important lexical resources for English is the British Parliament (Hansard) corpus[2] which contains 7.6 million speeches between 1803-2005, created by Anderson & Alexander [2011]. It is based on the TIME corpus, containing100 million words of text in about 275000 articles from TIME magazine, issued between 1923-2006.

Up to the 16th century almost all scientific writing in Europe was conducted in Latin. For instance, in [Onelli *et al.*, 2006] it is presented the construction of a diachronic corpus comprising written Italian texts between 1861 and 1945 from the DiaCORIS project, started in 2004. This corpus is based on other three lexical resources, freely available in online space for research on old Italian: OVI (*Opera del Vocabolario Italiano*). It includes texts in prose and poetry, written from the beginning of the XIII century to 1375 (about 19 million words). It is very useful for the automatic generation of concordances and word lists and BIVIO (*Biblioteca Virtuale On-Line*) that includes 200 literary texts by about 60 authors of the XVXVII centuries, and LIZ (Letteratura italiana Zanichelli) with 1000 literary texts from the XIII to the XX-

[1] An interactive web application: http://books.google.com/ngrams
[2] http://www.hansard-corpus.org/texts.asp

century, that contains a complex query program of LIZ able to retrieve data information on a single text or any combination of these texts.

Furthermore, in [Sánchez-Marco *et al*., 2010] a diachronic corpus for Spanish integrated into IAC (*Corpora Interface Access*) is presented. We have to consider the creation and improvement of diachronic language resources (POS-tagger), where are included over 20 million words of fiction and non-fiction texts from the 12th until the 16th century [see Corfis *et al.*, 1997; Herrera & González de Fauve, 1997; Kasten *et al.*, 1997; Nitti & Kasten, 1997; O'Neill, 1999; Sánchez-Marco *et al*., 2003] compiled, transcribed and edited by the Hispanic Seminary of Medieval Studies (HSMS) and other electronic corpora such as: CORDE[3] or Corpus del Español[4]. The authors mention the manual annotation with morphosyntactic information or automatic tagging that was, then, checked and followed by human correction.

We also identified the interest for developing some large historical linguistic resources for German. In [Lüdeling *et al.*, 2005] the DeutschDiachronDigital[5] is described, being a large corpus of old German texts, annotated linguistically (at lemma, POS, inflectional morphology, syntax trees, etc.) in order to implement a language model. The starting point was the old Saxon texts from the 9th century up to modern German at the end of the 19th century. For Old High German to Chadic and West African languages in [Chiarcos *et al.*, 2008; Zeldes *et al.*, 2009], a flexible web-based corpus architecture is described for search and visualization of multi-layer linguistic corpora (various annotations types), called ANNIS (ANNotation of Information Structure). It offers support for diverse annotations according to multiple annotation schemes that interact with all levels of grammar (phonology, prosody, morphology and syntax, semantics and pragmatics).

For Slovenian, the annotation process of the historic corpus is described (19th century) in [Kenter *et al.*, 2012] and presented in [Erjavec, 2012]. It contains 1.000 sampled pages from over 80 texts, written between 1750-1900, for the most part. The annotation consists in modern-day word form equivalent, lemma, POS tag, and optional gloss, using the Web-based editor, CoBaLT (Corpus-Based Lexicon Tool). In addition, in [Rissanen, 2008], the utility of one of the most known annotated diachronic corpora is presented, the treebanks built from the Helsinki corpus. The author presents the interplay of extralinguistic and linguistic factors in language change.

Three years ago, we started to develop a historical corpus for Romanian, based on a large newspapers collection from Romania and Bessarabia. This work begins from 1829, considered

[3] http://www.rae.es
[4] http://www.corpusdelespanol.org
[5] http://www.deutschdiachrondigital.de/

the reference year for the Romanian print media (see more in Chapter 6). The main problem that we faced was that in the 18th and 19th centuries in written press both Latin and Cyrillic scripts were accepted. This is one of the reasons why the development of this historical corpus has been slowed. We had to include a new method for language processing, the transliteration from Cyrillic to Latin. Now, there are two research groups, one from Romania and one from Bessarabia, who have joined the forces to do such work [Boian *et al.*, 2014; Petic & Gîfu, 2014; Gîfu, 2015a; Gîfu, 2015b]. The final goal is to analyze the differences between two cognate languages regardless over time.

For Romanian language some studies revealed the morphology diachronic reconstruction, using the quotations from eDTLR (Dictionary Thesaurus of the Romanian Language in electronic form) [Cristea *et al*., 2012; Gîfu, 2016b]. Furthermore, based on diachronic corpus for Romanian, with over 3 million words of text for now, the recovery process of old lemma is not finished yet.

In this context of the development of the historical corpora, it worth to remember, without detailing it, another important task: the creation of a parallel corpus of diachronic language. In general, such a corpus is formed by biblical texts, because the Bible is one of the earliest sizable coherent texts documented for many languages (especially in Europe). 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 example, for older Germanic languages [Sukhareva & Chiarcos, 2014].

2.2 Diachronic text classification

Text classification is the most popular task in NLP. It was and is still used also in diachronic studies. The authors of a major study [Szymanski & Lynch, 2015] propose a statistically method based on extracting n-gram features from each text - n-grams (Char), part-of-speech tag ngrams (POS), word n-grams (Word) and syntactic phrase-structure rule occurrences (Syn) -, in order to train a Naïve Bayes classifier ($Google_{nb}$) on date-labelled training data, using a diachronic corpus for English. Their corpus contains news articles from The Spectator[6] archive, published from 1700 until 2014. Additionally, they used a huge external source of data, Google Books Syntactic N-Grams (GSN) database [Goldberg & Orwant, 2013], focusing only on sample of 1 million

[6] http://www.spectator.co.uk/

English-language books dating from 1520 to 2008.

The previous approach is similar with [Argamon-Engelson *et al.*, 1998] in which is described a stylistic text classification (POS and word n-gram), according to their source, with statistically-detected stylistic variation. Moreover, Szymanski and Lynch used the syntactic phrase-structure rules encountered in [Feng *et al.*, 2012], used as stylometric features - driven from CFG (Context Free Grammar) parse trees - for deception detection.

2.3 Languages similarities using machine learning

The comparative study of the Romance languages (French, Italian, Portuguese, Spanish, and Romanian) exposes the presence of many similarities especially in diachrony as illustrated in [Densuianu, 1902:23]. The main problem is that the Romanian language has had a different evolution than other Romance languages, because of its isolation. Eventually, "the language is made, but its accomplishment process is a historical work, not a daily one." [Coșeriu, 1997:246]. A major contribution in diachronic syntax in general, focusing on comparative and diachronic Romance languages, and the diachronic syntax of old Romanian language in particular, is highlighted in [Pană Dindelegan *et al.*, 2015]. The authors describe and analyze the history of the Romanian language from the perspective of the diachronic syntax.

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 & Cristianini, 2010].

The development and use of software for natural language processing (NLP) highlight the defining aspects of the Romanian print press (morphological and syntactic analysis, semantic analysis and, more recently, pragmatic analysis) that have many similarities with that from Bessarabia on the time axis that we have chosen.

The differences between the lexicons of the two languages [Gîfu & Simionescu, 2016] then for four [Gîfu, 2016c] versions of a language (idioms), from the perspective of transformation patterns were analyzed using **Maximum Entropy text classifier**. It has been successfully applied to diachrony and it is commonly used to develop a machine learning algorithm (see Chapters 6). All the substring replacement operations (referred as "REP") are

classified and extracted from the known correlations list, based on the character-level context in which they are applied in the source word. Based on these REPs, a set of fictive/candidate words are generated, each having a trust score attached. If a candidate word is found in the destination lexicon (DEX-online), the two words are marked as a corresponding pair.

In [Szymanski & Lynch, 2015] is used the SVM (Support Vector Machine) classifier. The authors performed different experiments with the Weka implementation of ordinal regression [Frank & Hall, 2001], using a SVM base classifier. However, the results showed a lower accuracy than the standard SVM classifier. Because of that, the authors preferred a standard multi-class SVM classifier for all of our evaluations and predictions. The final results reveal that a stylistic classification approach can predict exact data when a written text belongs to the sample category.

One of the studies that will underpins the language similarity module of this PhD thesis belongs to Popescu & Strapparava SVM, which describes the diachronic text evaluation task. [Popescu & Strapparava, 2015]. After their prediction of the epochs when there are important linguistic differences between the written press between 1700 and 2010, they used a number of statistical methods concerned by dating a text as a stylistic classification problem. For this task, they trained a multi-class SVM model using a set of stylistic features (character frequency, POS-tag, word, n-grams, and syntactic phrase-structure rule occurrences) extracted from their corpus.

To provide a complex picture for language similarity, two models **Latent Semantic Analysis (LSA)** and **Latent Dirichlet Allocation (LDA)** are important. Taking into consideration our studies, theybest reflect the differences in writing style, ranging from sentence and paragraph structure.

To study of topical changes across periods of time in [Blei & Lafferty, 2007; Wijaya & Yeniterzi, 2011] used the LDA model. Related to epoch identification, Wang & McCallum [Wang & McCallum, 2006] develop time-specific topic models to a time stamp prediction task.

Yuan and Qian [Yuan & Qian, 2015] used LDA model in cross-language text similarity computation to obtain Tibetan-Chinese comparable corpora. Although LDA was proved to be useful, some factors affect the correction of cross-language text matching (the accuracy of Tibetan word segmentation, incomplete Tibetan-Chinese dictionary the particle size of Tibetan and Chinese word segmentation).

For this PhD research, we use LDA and LSA models to compare the writing styles in the Romanian language. The comparison is based on the period from which the texts originate (1941-1991 vs. since 1992), on one hand, and on the region (Romanian vs. Bessarabia), on the other hand. We are aware that the present data are still insufficient, which is why these models

were used only this time due to important differences in word structure (old, absolute forms not present in newer vocabularies), as well as to completely different topics of interest and corresponding concepts. [Gîfu *et al.*, 2016]. In order to keep the categories balanced, we have selected around 37 contemporary journal numbers from the four categories corresponding to each geographical area (Moldavia, Transylvania, Wallachia and Bessarabia).

2.4 Languages similarities using machine translation

The similarities between languages are interesting for historical and comparative linguistics, as well as for machine translation and language acquisition. There are works that focus on the possibility of obtaining a better quality in translation using simple methods for very closely related languages.

In [Hajič *et al.*, 2000] is described the Česílko system, affirming that the translation memory (TM) can be used as independent language. The authors present a method of a "triangular" translation for a set of cognate languages (here, Czech and Slovak), like in our case, through a pivot language using the human translation in parallel with machine translation. Furthermore, they demonstrated that the proposed method can be adapted for a new language (here, Polish language).

In [Koppel & Ordan, 2011] the impact of the distance between languages on the translation product is presented. The authors conclude that it is directly correlated with the ability to distinguish translations from a given source language from non-translated text.

In [Scannell, 2014] a statistical model for translation from Scottish Gaelic to Irish is presented, considered closely-related languages in order to discover the orthographical differences between them. The authors consider that a machine translation from Scottish Gaelic to Irish can be used effectively in this problem. For this task, they applied a statistical model, namely the standardization of historical Irish texts, used before in [Scannell, 2006]. It represents the conditional probability of some source language word (Scottish Gaelic) corresponding to a given target language word (Irish).

2.5 Language similarity metric

Which metric is best suited for the language similarity task? 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 in many works. For instance, in [Frunză *et al.*, 2005] a multilingual text processing tool is described, called BALIE1. It is designed to support information extraction for English, French, German, Spanish, and Romanian. The authors demonstrate how they adapted this application for the Romanian language (known as RO-BALIE system), starting with four functionalities: language identification, segmentation, tokenization, and POS tagging.

At the same year, in [Dinu & Dinu, 2005] the syllabic similarities between Romance languages (Latin, Romanian, Italian, Spanish, Catalan, French and Portuguese) in different epochs are presented, using the vocabularies developed by Sala [Sala, 1988]. The authors constructed, for each vocabulary, a classification of syllables. Then, they applied the rank distance described in [Dinu, 2003] between all pairs in order to compare each Romance language with the others.

The studies focused on language similarity that included also the Romanian language, followed then the determination of the orthographic similarity, described in [Ciobanu & Dinu, 2014]. The authors proposed three metrics. For measuring the distance between two words, they used LCSR (the longest common subsequence ration) described in [Melamed, 1995], a normalized version of LD (Levenshtein distance), dividing the minimum number of operations accounted by LD [Levenshtein, 1965] by the length of the longer string, and the RD (Rank distance), mentioned before. The authors observed the etymology (lexical resources in a foreign language) and cognates (identifying words in different languages with the same etymology, including a common origin) between Romanian and related languages in three historical epochs: 1643-1743, 1870-1889, and nowadays.

Moreover, the authors mentioned that they were inspired from [Rama & Borin, 2014], probably the most comprehensive study that presents 20 different similarity (or distance) measures in order to solve the problem of languages relatedness classification. The authors describe each metric, considering the cognate identification and tree inference tasks, closely related in historical linguistics.

2.6 Epoch detection

There is a major interest in building diachronic systems, the epoch's detection on the time axis, when we analyze the evolution of language or we want to detect a certain period of belonging for a particular document, being actual tasks in NLP.

Regarding epoch detection starting with the semantic changes over time, we remind a study [Wang & McCallum, 2006] focused on the analysis of topics over time. The paper focuses rather on fixed topics, which are expressed by frozen compounds, such as "Mexican war", "CVS operation". The authors determine how these topics evolve during the years. However, because the scope of their paper is not global, the corpus used comes from 19 months of personal emails. It is hard to see how this method could be generalized.

A similar approach is described in [Wang *et al.*, 2008]. The authors use LDA to facilitate the search into large corpora by automatically organizing them. Furthermore, in [Mihalcea & Năstase, 2012] it is shown how to detect automatically periods of time as epochs, considering semantic disambiguization. The authors used texts from Google Books and their approach is based on the fact that words change their neighbors over time. Furthermore, starting from this study in the SemEval 2015, for the task 7: Diachornic Text Evaluation, we found works that remain in the same research area.

Thus, in [Tan & Ordan, 2015] the primary concern remains the automatic epochs detection of the historical text snippets based on adapting a web crawler to extract data from the original source. The aim of this study is to determine the temporal information of the text snippets from their URLs, using machine learning classifiers for situations where test data is considered unknown.

In addition, in [Popescu & Strapparava, 2015] a large corpus is described, containing a collection of newspapers published between 1700 and 2010. The main goal is to automatically identify the period of time of which a particular written text belongs, using three systems. First system, called AMBRA (*Anachronism Modelling by Ranking*), is based on the learning-to-rank framework using pairwise comparisons. It is a method proposed for temporal text modelling by [Niculae *et al.*, 2014] which described how the statistical methods may be used in predicting the epoch to which the analyzed documents belong. Second system uses the entities names presented in text and their link to data from Wikipedia to learn the epoch they belong. Third system is based on the use of language features in combination with machine learning (SVM), using Google NGrams.

This AMBRA system is, also, described in [Zampieri *et al*., 2015] where the authors incorporate the label of uncertainty by assigning an interval within which the document was written.

Moreover, one of the works that formed the basis of our study is found in [Popescu & Strapparava, 2013]. The authors use two statistical methods (SVM and LSTM) for detecting epochs during 3 centuries, LSTM (*Long Short-Term Memory*). It is a novelty for Romanian

language.

2.7 Conclusions

Reading the studies published by our predecessors helped us to better perceive the differences occurring in the Romanian language, over time in different geographical areas. As we emphasized, given that the evolution of a language involves changing and adapting to the needs of speakers, our diachronic system is developed based on linguistic analysis highlighted by a new methodology proposed in this PhD thesis (Chapter 7).

To investigate and to get the feature importance, we consider several techniques. Moreover, is hard to decide what feature selection is good for language similarity. With other words, a clear conclusion could not be drawn regarding the most appropriate for language similarity task. To obtain good results, we consider that the corpora plays a huge role in corpus-based language processing methods (e.g. POS-tagging). Furthermore, we believe that by combining these approaches offers a comprehensive perspective on language similarity.

3. INTRODUCTION TO A LEXICAL SEMANTICS STUDY

This PhD thesis uses a particular formalism of the field of lexical semantics, namely statistical methods of linguistics introspection, to investigate the written Romanian language in use on different geographical areas and covering the period from the second decade of the 19th century till our days. This type of investigation is usual in classical linguistics, but we, basically, are using computational models and methods with applications in linguistics.

In this chapter we shortly overview the main contributions of lexical semantics in text analysis from a linguistics and computational linguistics perspective. Directly or collaterally, many of the presented studies will form the basis of a contrastive analysis of the collection of publications that we describe in Chapter 6. Our goal is to highlight the evolution of the language in time, in two geographical areas, Romania and Bessarabia, considering both historical and language factors.

There is a long philosophical tradition, by going at least from Parmenides and Socrates, which attests the importance of words in deciphering the meaning[7]. In the 20th century, Frege[8] [Frege, 1892/1952], Carnap[9] [Carnap, 1972], Wittgenstein[10] [Wittgenstein, 1991] and other authors were centered on the nature of meanings, the way words are used in different conditions, the relationship between words and their meanings (see Chapter 4), etc.

For James Higginbotham [1989:470]: "What is crucial to the design of language is not some distinction, however drawn, between properly semantic information, on the one hand, and

[7] There are known the concerns of Socrates about the modalities of defining the natural world, without to succed to overcome the arguments of Parmenides which refers the distinction between appearance and opinion. Parmenides was the first who distinguished between truth and appearance, the logic of a vision and the logic of the words. He also brought in discussion the objective knowledge issue. It is why Socrates was focused on discovering facts about the natural world, relying on the rigorous analysis of some afirmations which relate to facts.

[8] Frege makes a distinction between sense and signification and between *meaning* and *significance*, specifying that the sense in linguistic expression is its meaning and the significance is itself subject designated (concrete or abstract). He mentioned in his essay "On Sense and Meaning" that "a proper name (word, sign, sign combination, expression) expresses its sense, means or designates its meaning". [Frege, 1892/1952:61].

[9] Carnap improves the Frege's semantics method, the intensions and extensions method, seeking to rebuild the natural language after the exigencies of the logic and science. Moreover, he removes the "doubling" of objects in classes and attributes (properties).

[10] Wittgenstein distinguishes between sense, that represents the content of the descriptive expressions, and significance as assigned subject, having sense only the sentences that give state of things. The descriptive theory of the meaning, which the Austrian philosopher adopted here, obliges him to use the table method of the truth functions, when he needs to determine the meaning of the complex sentences. Later, he abandoned the descriptive theory from Tractatus in the favor of the pragmatic language theory, focused on *language games* [Wittgenstein, 1993], like speech acts, defined by Austin.

empirical or collateral information, on the other, but rather the distinction between information that has systematic grammatical effects and information that does not".

In fact, the grammar is the basis of the linguistic concerns, mainly focusing on syntax and its relation with the morphology and semantics. During this PhD thesis, we are concerned to bring linguistics evidence (the lexical fund diversification, the meanings associated with the Romanian lexicon, the grammatical structure, etc.) in tracing the Romanian language evolution.

3.1 What is and what is not lexical semantics?

Lexical semantics (LS) or words' semantics is *the study of the meaning of words*. In the applicative scope, LS focuses on the lexical study (content) of the meaning of a word [Jurafsky & Martin, 2000], as opposed to the grammatical meanings of the words (functions). Moreover, lexical semantics is concerned about why the words mean what they mean, how they are represented in the mind of speakers and how they are used in a text. In other words, the lexical semantics has a deeply interdisciplinary character, being at the crossroads with at least cognitive science, computer science, and philosophy (mentioned above).

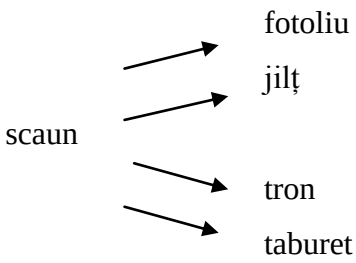
Lexical semantics, within linguistics, have an intersection with lexicology, but the last one studies merely the vocabulary of a certain language [Cruse, 2000]. In addition, it provides the basis for various applied research fields, such as acquisition, processing and language learning [Robinson & Ellis, 2008; Gullberg & Indefrey, 2010; Pütz & Sicola, 2010], lexicography [Béjoint, 2010], and computational linguistics [Pustejovsky & Bergler, 1992; Saint-Dizier & Viegas, 1995; Klavans, 1995; Asher & Lascarides, 2003]. If a native Romanian speaker learns a foreign language starting by accumulating new lexical knowledge.

All assertions previously mentioned motivate our research effort in this PhD thesis from the perspective of the words meaning in time. We are interested to find an answer to the question: *how have changed the lexical meaning of words in time?*

The term "meaning" was proposed by Saussure [de Saussure, 1916], to signify an ensemble of significant units (*semes*), existing independently of the utterance. The meaning of a word reported to the meaning of the words from the specific system to which it belongs [Firth, 1962] is defined by the opposition with the meanings of other words of the same language, where every meaning is a component element (bellow, an example). This idea was adopted by Guiraud: the meaning of each sign is encircled from all sides and delimitated by the meaning of the signs which are around it. [Guiraud, 1958].

For instance:

The meaning of *scaun*[11], in the structural lexical semantics, is defined in opposition to other words of the same lexical-semantic paradigm, called, *scaunelor* → (En): chairs.



Over time, the way to perceive the meaning has become more and more complex, because the researchers have attributed to almost anything the quality of bearer of meaning. We think about the constituent parts of the word, the melodic structure of the communication, the relationship with objects, the processes, etc. Besides the main meanings of language - lexical and grammar - the stylistics adds the emotional-expressive meaning. It is what some linguistics called the "semantic hypertrophy" when they refer to meaning [Ogden & Richards, 1923; Schaff, 1966].

Historically, the traditional lexical semantics is known as linguistic semantics, and is mainly descriptive. It studies the causes of meaning changes or meaning shifts that occur inside a language [Șăineanu, 1887; Bréal, 1897; Sperber, 1922; Ullman, 1952; Guiraud, 1964; Meillet, 1975]. Along the decades 50s - 60s of the previous century, Baldinger [1952; 1960], Ullman [1962] and many others have put the foundations of diachronic semantics.

From the few works mentioned, we can say that *lexical semantics* can be understood as:

(1) *semantic morphology* (e.g. the study of the meaning of compounds words, whether and how can the meaning of the compound be assembled out of the meanings of its constituents), often considered a separate domain of what is known to be *lexical semantics*.

(2) *the study of the meaning of word*, which includes aspects that are part of the large domain of lexicology (e.g., lexical variation beyond linguistic diversity, phraseology, loans, etymology, etc.).

[11] After identifying the common seme specific to the constitutive terms of this lexical-semantic paradigm, the specific seme of the unit of meaning analyzed will be identified: /object to sit/, / for a single person/, / with or without backrest/, /with or without arms/.

When we refer to the lexical meaning[12] our attention stops on all the essential features that the speakers assign to a word.

3.2 About theories of lexical semantics

Depending on the tradition of the school the authors belong to and the period of elaboration, lexical semantics theories give more or less attention to the representation of the lexical level [Paradis, 2012]. Lexicographical approaches, for instance, are focused on the need to understand how the entries should be written in a dictionary. Jackson's workbook [Jackson, 1988] carefully illustrates a descriptive analysis of words, being concerned with the meaning of the words and the internal relations between them. In [Hudson, 1986; 1995] the organisation of the lexicon has been developed on the basis of the lexical semantic concepts with a brief theoretical description, Jackson & Zé Amvela [2000], Singleton [2000], Hanks [2007] and many others examine basic semantic problems, such as polysemy.

A number of computational approaches to the representation of polysemy can be found in [Martin, 1996; Copestake & Briscoe, 1995; Fellbaum, 1998, etc.]. The last author is known for her seminal work on WordNet, perhaps the most salient lexical resource used in NLP researches (see more in the next chapter). The meaning of language is strongly focused on lexical issues: the subtlety and diversification of meaning [Cruse, 2000] or the diachronic semantics [Geeraerts, 2010].

Lexical semantics is generally addressed from non-formal positions. Two general types of approaches can be analyzed – the componential and the schematic image -, which we compare below. Goddard's semantic analysis [Goddard, 1998] describes the meaning of words and sentences in order to reveal the bridge between language and culture. In this context, previous general works on semantics should be mentioned. In [Bréal, 1900] the rhetorical intention of the metaphors in everyday language is well described, considered a ubiquitous tool of linguistics innovation. Sperber [Sperber, 1922] claimed a similar theory, noting the enlargement of the semantic areas of words, using metaphors. Also, a few modern studies of the language change [Löbner, 2002; Saeed, 2003] are based on the Stern's work [Stern, 1931] that created a classical study of semantic change.

Here are some relevant theories, seen from the two mentioned perspectives:

[12] With respect to their capacity to possess or not a lexical meaning, words are called *significant* (morphological categories such as: noun, adjective, numeral, pronoun, verb and adverb) and *insignificant* (article, preposition, conjunction). Situations may arise when certain significant words lose their lexical meaning, phenomenon known as "desemantization" (e.g. auxiliary verbs).

- *componentials*: conceptual semantics, a framework for semantic analysis [Jackendoff, 1983; Jackendoff, 2002], generative lexicon focused on the distributed nature of compositionality in natural language [Pustejovsky, 1995], natural semantic metalanguage, "a decompositional system of meaning representation based on empirically established universal semantic primes" [Goddard, 2009; Goddard, 1998; Wierzbicka, 1996];
- *schematic*: theory of cognitive semantics [Evans & Green, 2006; Croft & Cruse, 2004], semantic framework – FrameNet[13] [Filmore, 1976].

Back to the componential (or *semic*) theories, they are based on a language system similar to the representation of the meaning in a grammar that involves a restricted number of "primitive" symbols. A classic example is the model introduced by Katz & Fodor [1963]. Regarding the schematic theories, they emphasize that the meaning of a word must be addressed in complex conceptual structures. These often rely on a form of representation oriented on an image, such as image schemes [Lakoff, 1987; Johnson, 1987]. In other words, the componential approaches are associated with the generative linguistics, while the schematic ones with cognitive linguistics. However, between the two categories exist many differences. Note, that the 80s are characterized by an explosion of grammatical theories[14], all having the same origin – generative linguistics.

Lexical semantics starts to emerge as a research discipline in Aristotle's time, but it was shaped especially in the 19th century. Then, a number of researchers with a large interest in historical texts and human culture have addressed this problem, unfortunately without much coherence. Lexical semantics at that time had, merely, a historical and philological orientation. It was mainly concerned about etymology and to understand how the meanings of the words change in a language over time. The meanings of the words were regarded as mental constructions (thoughts) and, as such, the process of meaning shifts or changes being the result of psychological processes [Geeraerts, 2010].

Gradually, in the 20th century, the followers of the structuralism school, associated with the work of Ferdinand de Saussure [1916/1959], move the perspective from diachronicity towards synchronicity. According to the structuralists, the language is an intralinguistic

[13] http://framenet.icsi.berkeley.edu

[14] In the context of the linguistics, the consecrated theories are: *Theory of Government and Anaphoric Dependencies* (known as *The Principals and Parameters Theory* and revised in *The Minimalist Program)* [Chomsky 1981, 1995]; *lexical-functional grammar (LFG)* [Bresnan 1982], *Generalized Phrase Structure Grammar* [Gazdar *et al.*, 1985], *Categorical Grammar* [Oehrle *et al.*, 1988] and *Head-driven phrase structure grammar (HPSG)* [Pollard & Sag 1987, 1994].

autonomous system of relationships between words[15], organized on lexical fields [Lehrer, 1974; Cruse, 1986].

For instance: *lung* (En: *long*) has a meaning through its relationship with *scurt* (En: *short*), which is an antonym.

Basically, the structuralists have made a distinction between paradigmatic and syntagmatic relations. The last ones are linear relationships formed between words in a Cruse's type sentence [Cruse, 1986:6]. A Cruse sentence claims that we cannot imagine the meaning of a word as affinity models with all the other words of the language with which it is capable of semantic relations contrasting in grammatical contexts. There are two types of affinity: syntagmatic and paradigmatic. Cruse affirms that the syntagmatic relations are "discourse cohesion devices which add necessary informational redundancy to the message" [Cruse, 1986:86].

For instance: *câine* (En. *dog*) and *latră* (En. *bark*)
Câinele latră. → (En) *The dog barks*.

Indeed, a syntagmatic affinity is determined by the ability to be part of an association in an expression.

A syntagmatic disagreement is revealed by an anomaly that does not violate the grammatical constraints.

For instance: *lei* (En. *lions*) and *zboară* (En. *fly*)
Leii zboară. → (En) *Lions are flying*).

The ideas of the syntagmatic approach, of what strings of words mean, of collocations and models of co-occurrences, were developed within the London school [Firth, 1957] and the Birmingham school [Sinclair, 1987]. The syntagmatic approach preserves in the language use all the meanings of a word or the communicative meaning of language. It opened new trends in linguistics. We refer here to those lexical semantics approaches that state the importance of contextual factors and real language usage in the description of the meaning (e.g. cognitive linguistics).

Towards the end of the 20th century, there was a reaction against the structuralist view.

[15] According to G. A. Miller [Miller, 1985], lexical relations can be divided into the following classes: homonymy, polysemy, synonymy, antonymy, hyponymy, meronymy, etc.

The language became to be seen as a system of relationships between words. It returned to the opinion that the meanings of words can be considered psychological entities located in people's minds, rather than relationships between words. This renewed interest has occurred while improving investigation methods benefiting from the technical progress and computerization in research. The structuralism was replaced with two branches totally different of conceptual approaches in lexical meaning: *generative* and *cognitive*.

In the article *The structure of semantic theory* [Katz & Fodor, 1963], the meaning is described as a part of a formal grammar, in terms of the meaning of components.

For instance: `femeie = {+ om, + feminin, + adult}`

The main purpose of this theory in generative lexical semantics was to develop a logical formalism to be used either for the meaning of words' deconstruction, as developed by Jackendoff [1983, 1987, 1990, 2002] and Pinker [1989], or to build the lexical meaning, as in the generative lexicon [Pustejovsky, 1991; 1995], [Pustejovsky & Boguraev, 1993].

Ray Jackendoff aims to restore with formalized means how people express what they understand from the sense of occurrences of a word. Jackendoff's theory [Jackendoff, 2011] explains what is going on in the minds of people who use the language. Knowledge about the language of a speaker is a set of mental structures and relationships between structures plus the ability to combine these stored elements. This theory is represented graphically in Figure 1.1:

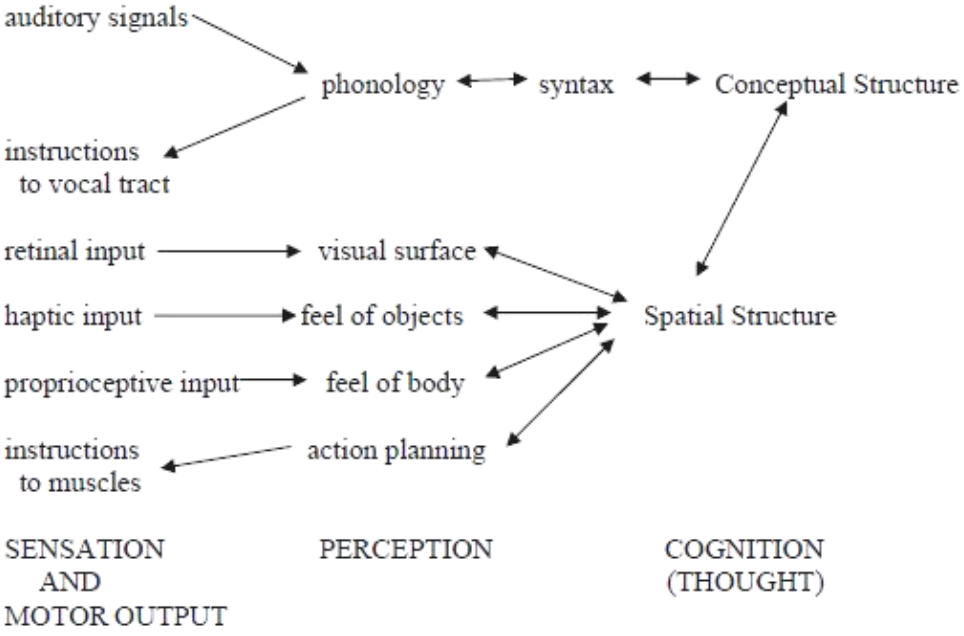


Figure 1.1 Architecture of the mind (reproduced from [Jackendoff, 2002])

According to Jackendoff, the language can refer to ontological referential and aspectual categories, as follows:

(1) Te rog, poți ridica asta [*arătând*]? → (En) Could you raise that [*pointing*], please?

[reference to object]

(2) Poți pune haina acolo [*arătând*]. → (En) You can put your coat there [*pointing*].

[reference to location]

(3) Ea s-a dus în direcția aceea [*arătând*]. → (En) She went in that direction [*pointing*].

[reference to direction]

(4) Ați putea să mă ajutați să termin asta [*demonstrând*] → (En) Could you help me with finishing this [*demonstrating*]?

[reference to action]

(5) Ai auzit asta? → (En) Did you hear that?

[reference to sound]

(6) Mai bine că nu s-a întâmplat asta [*arătând*] în biroul meu. → (En) Better that [*pointing*] never happened in my office!

[reference to event]

(7) Am prins un soarece atât [*demonstrând*] de lung. → (En) I caught a mouse that [*demonstrating*] long.

[reference to distance]

(8) La grevă au fost atâția [gesticulând] oameni. → (En) At strike were this many [gesturing] people.

[reference to amount]

(9) Invitatul nostru va intra pe scenă chiar acum… [applaudând] → (En) Our guest will enter on scene right now… [*clapping*]!

[reference to time]

(10) Poți să te porți așa [*demonstrând*]?→ (En) Can you behave like this [demonstrating] ?

[reference to manner]

The activity was conducted recently on formalizing of the meanings in context, as a function of organizational rhetoric. [Lascarides & Asher, 2007]. In the recent models, we noted that decomposition of simple traits was replaced by models that take into account the implicit structure and social frameworks. Today, the language users know many things that are relevant

for the linguistic interpretation. The ultimate purpose of these formalizations is to be useful for different types of computational implementation in information technology and computer science. Similarly, in generative linguistics, the lexical meaning in the cognitive linguisticsimplies that the meanings of linguistic expressions are mental entities.

However, the two frames are different. We refer to the fundamental assumptions about what is the meaning in language, how the words relate to meanings and how the meanings are described and explained. [Paradis, 2003]. We use the words closely related to the inter-subjective conditions in an act of social communication. There is no strict dichotomy between linguistic and encyclopedic meanings. We could say that the lexical meaning in cognitive linguistics is encyclopedic, if we consider every aspect of the modulation of the contextual sense. The contextual sense is relevant for a specific task, integrating semantics, pragmatics and, in fact, also grammar.

Note that in the contemporary linguistics, the cognitive lexical semantics is the most popular, both in terms of publications and number of people involved in the lexical semantics research. [Geeraerts, 2010].

Why? Because there is a direct mapping of the words and the phrases of the conceptual structure. The words function is to evoke conceptual models in the cognitive system. It is no longer a novelty for anyone that the language is an integral part of human knowledge.

There is no linguistic algorithmic intermediary level between the language expressions and their meanings. The cognitive approach of a meaning is based on understanding the language context. [Langacker, 2000; Cruse, 2002; Tomasello, 2003]. Speakers and listeners are intentional creatures. How we express is functional motivated; we want to get our message through our interlocutors and to negotiate optimally meanings in the communication with other people. Unfortunately, there are no stable meanings of the words. We could say that the meanings of the words are dynamic and interpretable. Therefore, they are context-sensitive.

Thanks to technical innovation as a form of the increased computer capacity and its performance, and of the improving the experimental equipment, a lot of empirical progress has been achieved lately. The contemporary research in the lexical semantics is the widespread use of the methodologies based on corpus and the language technology, often called human language technology (HLT). Basically, the use of large databases and the Web have revolutionized the investigation possibilities of the patterns of use a real language. We remark that the data types and linguistics culture are evolving permanently, based on probabilistic ideas. Besides all these views an important role in lexical semantics research plays different types of psycho- and neurolinguistic experiments. They are used to describe and explain the meanings of the words in

order to establish a connection between language and cognition, language in the minds and brains of people.

3.3 The nature of lexical semantics

If most of the semantic theories published before the 90's refer to how meanings are constituted in phrases and sentences, and the similarities and differences that exist between the semantics of natural language sentences and phrases and semantic judgments in logical languages, lexical semantics instead is that part of semantics that only refers to word meaning. [de Saussure, 1922]. As we mentioned above, Saussure argues that the meaning of a word is given by the differences between signifiers: syntagmatic (related to position) presented on the horizontal axis and paradigmatic (related to substitution) on the vertical axis.

Moreover, the syntagmatic relations are variants of combinations and those paradigmatic involve differences. In natural language, there are grammatical paradigms such as verbs or nouns. We emphasize that the structural assumptions are also adopted by the computational linguistics.

For instance: WordNet [Miller *et al.*, 1990] with Romanian version, RoWN (*Romanian WordNet*) that we will describe in the next chapter. They are dictionaries that describe the meaning of a word through semantic networks.

However, we highlight that the main limitation of the "radical" view of the structuralism is that it is almost impossible to describe all the possible associations between all the natural language terms. It is an impressively large number of concepts and semantic relations between them.

The literature contains many studies on the lexical semantics topic, signed by: Carter [1987], Hurford & Heasly [1983], Palmer [1982], Crystal [1987] with large accessibility. A special attention we should also give the works signed by: Fromkin & Rofman [1988:205-250], Jackson [1988], Leech *et al.* [2009], Kempson [1977], Cruse [1986], Lyons [1977], ș.a.m.d. The works of the last two linguists, Lyons [1977] and Cruse [1986] are classic linguistics texts on lexical semantics where the lexical semantics nature has changed significantly.

Cruse [1986] established in a principled and disciplined manner the generalized descriptive facts about lexical relations that any formal theory of semantics will have to include. Among the topics covered, in depth are lexical ambiguity, synonymy, and hierarchical relationships, such as hyponymy and meronymy, and different types of opposition (see Chapter

4). Combinatorial characteristics of the word in context are constrained not only by understandable characteristics, but also by grammatical properties. Features on the semantic grammar strengthen them, but can be arbitrarily semantic. [Cruse, 1986:1-4].

If during and after the 80s, the grammatical theories have focused more on the lexical level, today there is a tendency to present the lexical semantics with a limited accent to the polysemy / homonymy relations "NYM"[16] (synonyms, antonyms, acronyms etc.). The lexical meaning was and is the central element in the computational linguistics. One of the most popular Romanian online lexical resources that we use in the contrastive analysis (see Chapter 7) is DEX-online[17] which also includes lexical guides: dictionary of synonyms and dictionary of antonyms. [Seche, 1997; 2002].

Here some examples:

The acronym *UE* (Uniunea Europeană) → (En) *EU* (European Union)

bolșevic with the synonim *comunist* → (En) *communist*.

propriu with the antonyms: *comun, impropriu, figurat, necaracteristic.* → (En) *proper*: *common, own, figurative, uncharacteristically*).

3.4 Levels of language facts

After this incursion over the most important aspects that are the subject of lexical semantics, in this section we mention the language meanings facts, without developing them, useful from the perspective of computational linguistics. Semantic study can be conducted on the following levels:

- *lexical level*, at which the sign is presented as having an intensity, a definition of meaning, and an extension, a lot of real objects which match to that definition.

- *propositional level*, at which a judgment is issued through a combination of signs. Its intension is a reunion of signs components intensity and the extension is the truth of the proposition. In a Boolean system, it may be 1 for True and 0 for False, but there are other logical systems with multiple truth values.

- *sentence level*, at which calculations can be performed on sentence components and their truth values.

- *discourse level* (termed texts or narratives) is the next level in a hierarchy of

[16] NYM means name or word (http://www.fun-with-words.com/nym_words.html).
[17] www.dexonline.ro

morphemes, clauses and sentences [Harris, 1952].

Actually, all these levels of analysis can be said to be a Natural Language Processing (NLP)-based system. With other words, computational linguistics research is gradually evolving from lexical semantics to compositional semantics and narrative understanding.

3.5 Conclusions

The previous remarks attest that the structural lexical semantics has radically modernized the linguistic semantics and that the computational lexicography uses its principles andresults. Furthermore, the diachronic semantics has in attention reducing the diversity of the individual meaning changes to some analogous general trends in historical phonetics. The semantic evolution of the words is reduced to the meaning change, the widening and restricting its.

For instance: in the Transylvanian publication name *Foaia pentru Minte, Anima și Literatura*[18], the word *anima* changed its meaning over time.

The Latin term, *anima* (in Germain, *anemos*), has a parallel evolution [Niculescu, 2009] as follows:

The original meaning is „souffle, vent, exhalation, respiration" (see Quicherat – Daveluy, Chatelain dictionaries), alongside the meaning „principiu vital", „existence, âme" (cf. *constare hominem ex anima et corpore dicant*, to Pacovinus, tragedian, sec. II î.Ch.) and later „suflu al vieţii" (*animam extinguere* to Terentius). Starting with the 3rd century, we find the word *anima* to Christian authors, as Hieronymus, meaning exclusively „sufletul omului" (*dilige Dominum ex toto anima tua*, Hieronymus).

Finally, we can say that the evolution of a language, based on meaning changes, has an extralinguistic cause (e.g. changing the socio-historical relations) or a linguistic one (e.g. the popular etymology). The exploitation the lexical semantics domain in natural language processing allows us to combine rationally the theories of those two fields, which will be reflected in the contrastive analysis (Chapter 7).

[18] More about *Foaie pentru Minte, Anima și Literatura* in [Hangiu, 2004].

4. ROMANIAN LEXICAL RESOURCES

Computerization of a language is strictly related to lexical resources, supposed to be used by both linguists and computer scientists. In this section, we will refer to three types of lexical resources: monolingual tesauri, dictionaries[19] and corpora of written texts. Some examples are: a lexical thesaurus – RoWN, the Romanian WordNet, a dictionary in electronic form – eDTLR, the Dictionary Thesaurus of the Romanian Language, and a diachronic corpus encripted in the Latin script[20] and including texts originated in four geographical regions (Romanian Moldova, Transylvania, Wallachia and Bessarabia) and covering a period starting with the second decade of the 19th century till our days (detailed in Chapter 6). With respect to the usefulness of sources (not necessarily electronic) to the study of language, Vintilă-Rădulescu [2002] says: "the fundamental sources of information, conveniently stored (mostly in the traditional manner), are components of culture in the broadest sense. They are likely to be harnessed for the study of the Romanian language and for various related purposes, including those applicative, in the current computerized society." At the Institute "Iorgu Iordan", nowadays, the Institute "Iorgu Iordan – Al. Rosetti" were made the most popular Romanian language resources. [Avram *et al.*, 1999].

In the context of the computerized society, the language becomes technology, a subject of the technical investigation [Danzin, 1992; Tufiş & Slama-Cazacu, 1994]. The language technology requires specific methodologies for research / development, development, and adaptation of basic linguistic resources (dictionaries, thesauri, corpora, etc.) aligned with the practices, standards, and international recommendations. The alignment is essential for automatic processing programs of different languages (e.g. machine translation).

In this chapter, we present an inventory of the most known lexical resources in NLP. We

[19] Examples of dictionaries on paper form are: *Dicționarul limbii române* [DLR, 2006], with its old version known as *Dicționarul Academiei* (DA); *Mic dicționar academic*, Ed. Univers Enciclopedic, 2001, 2002, 2003; Coteanu *et al*. (coords.) [1975, 1996]; *Dicționarul explicativ al limbii române* (DEX), București, followed by a supplement and other versions. An example of an online dictionary: Dexonline at www.dexonline.ro.

[20] Considering that in the period 1940-1989 (excepting the war years), in Bessarabia publications in Latin script have been prohibited, all texts used in this contrastive analysis of a Bessarabian origin were transliterated from Cyrillic to Latin; details in [Boian *et al.*, 2014; Petic & Gîfu, 2014].

used it for the purpose of this PhD thesis (e.g. lexical ontologies as wordnets[21]), as well as other researchers in their contrastive diachronic studies on Romanian language, which determines the degree of similarity between cognate languages.

4.1 A historical perspective on Romanian lexical resources with a glimse into the future

If in the 90s, the Romanian language technologization was in its infancy [Tufiș, 1996], the constant effort of the Romanian researchers, linguists, and informaticians alike today is proving to be remarkable. In the early '60s, a group of young engineers and mathematicians from Timișoara, led by Erica Nistor Domonkos, was involved in a pilot project in the field of machine translation. Three years later, with the support of the scientist Grigore Moisil, a symbolic machine translation system capable to translate from English and French into Romanian is implemented [Domonkos, 1967].

All along the 60s and 70s, Romanian linguistics traverses a period of theoretical and methodological opening, influenced by a structuralist orientation (see Chapter 3). The Romanian linguistics school, led by professor Alexandru Rosetti, pursues theoretical investigations, applying the methodology of the structuralism current in the effort to offer a systematic description of the Romanian language. In parallel and supporting each other, the Romanian linguists school of structuralism intertwines with mathematical linguistics. The statistical approaches of Solomon Marcus, the creator of the Romanian school of mathematical linguistics [Marcus, 1970a], still have a major international recognition[22]. The scientist has important contributions in NLP (*Natural Language Processing*) by formulating mathematical models to analyze poetry, scientific language, and drama [Marcus, 1970b]. Marcu's ideas were continued by his disciples: Gh. Păun, on mathematics/computer science, and M. Dinu, on computational linguistics.

In the 80s, the interest for NLP grows, especially, when prof. Mihai Drăgănescu founded, in the Central Institute for Research in Computer Science, a new group of systematic research in Artificial Intelligence, including NLP. This is the period when the first NLP modern systems for Romanian language were implemented: BANDASEM [Dănăilă, 1980], SDLR [Tufiș, 1982],

[21] The monolingual WordNets have their synsets aligned to the equivalent synsets of Princeton WordNet (PWN). WordNet [Fellbaum, 1998] is one of the most known lexical resources for the English language. Computational linguistics cannot be efficient without a strong language engineering component.
[22] Google Scholer shows citations in 2016.

SIRLIN [Milea, 1982], SCD [Curteanu, 1983], IURES [Tufiș & Cristea, 1985], QUERNAL [Cristea & Tufiș, 1985], analysis and voice synthesis [Drăgănescu & Burileanu, 1986], SINAL [Toma & Săftoiu, 1987], PARADIGM [Tufiș, 1989; 1990], MORPHO [Tufiș & Dumitrescu, 1990] etc., with a broad applicability, including lexicographical researches. Technological progresses start to be visible in all areas of language processing, from the morpho-lexical level to the semantic and discourse level.

Although the research in theoretical linguistics, coordinated by renowned experts (Al. Rosetti, E. Vasiliu, I. Coteanu, M. Sala, S. Stati, M. Manoliu, A. Avram, M. Caragiu Marioţeanu, V. Guţu-Romalo, A. Bantaş, t. Slama-Cazacu, M. Avram, T. Cristea, A. Cornilescu, I. Dănăilă, M. Papahagi, S. Golopenţia, P. Miclău, I. Oprea, and many others) was performed, with very few exceptions, without adequate computational support, it represents the basis of applicative research for Romanian language.

After the 1990s, the boom in the IT sector reverberated also in the Balkans, Romania included, and, among other areas, has had a huge impact in the language technology, in general, and in the technologization of our language, in particular. The internet, the free circulation, the opened access of Romania in the European framework research programmes, and the scientific and academic exchanges contributed essentially to the development of language technology in our country. Basically, Romanian researchers started to be connected to the state of the art in NLP (recent methods and theories, newest methodologies, international priorities, etc.).

Gradually, a systematic concern regarding the acquisition of language resources for the Romanian language (machine readable dictionaries, thesauri, ontologies, word lists, language models, phonetic transcriptions, corpora, grammars, etc.) also started to appear.

As has been revealed rather recently in a thourogh study [Trandabăț *et al.*, 2012], the future of the Romanian language depends to a large extend on linguistic technologies, while these ones themselves are relying on language resources. Any evaluation of what will happen in the future should start from these premissed, to which the human component should be added. The progress depends so much on a clever and informed attitude of the political factors, that any estimation is difficult to make. What is sure by now is that Romania owns a number of qualified researchers and a number of centres in which the research in computational linguistics and natural language processing, with direct focus on the Romanian language, is active and closely connected to the mondial trends. To maintain and strenghten this orientation, still, a strategic nationally engaged vision should be created and promisses – be kept.

4.2 Romanian WordNet: RoWN

WordNet lexical ontologies are probably the most used lexical resources nowadays, in NLP applications with a considerable coverage: as reported by the Global WordNet Association[23] wordnets more than 50 languages. The Romanian WordNet (RoWN[24]) is one of the largest lexical ontology available today. Its roots should be searched some 15 years ago, in the Balkanet project [Tufiș *et al.*, 2004; Tufiș, 2008]. At present, RoWN makes use of 35 different relation types, some of them affecting no more than a few synsets. In total, there are 196.156 relations between 59.348 synsets[25]) [Dumitrescu, 2015], distributed among four parts of speech (POS): nouns (41063), verbs (10397), adjectives (4822) and adverbes (3066) (see Table 4.1). Like other wordnets, RoWN is designed to establish relations that involve all four POS types (our interest, in this PhD thesis, was focused on synonyms (e.g. love/like) and antonyms (e.g. love/hate).

PoS	Synsets	Literals	Unique Lit	NonLexicalised
Nouns	41063	56532	52009	1839
Verbs	10397	16484	14210	759
Adjectives	4822	8203	7407	79
Adverbs	3066	4019	3248	110
Total	59348	85238	75656	2787

Table 4.1 Statistics about synsets and literals in RoWN - [Barbu-Mititelu *et al.*, 2014]

The Romanian WordNet (RoWN) [Tufiș & Cristea, 2002; Tufiș, 2008; Tufiș *et al.*, 2004; Trandabăț *et al.*, 2012] is a semantic network built on the principles of Princeton WordNet (PWN), now at version 3.0 [Fellbaum, 1998], and aligned with it. Moreover, RoWN is aligned with the largest freely available ontology, SUMO&MILO[26], whose concept definitions are translated into Romanian [Tufiș, 2009]. RoWN is used in various applications developed for the Romanian language: questions answering systems [Ion *et al.*, 2008], word sense disambiguation [Ion & Tufiș, 2004], machine translation [Dumitrescu *et al.*, 2012], etc.

[23] http://globalwordnet.org/wordnet-annotated-corpora/
[24] The Romanian site can be found at the Research Institute for Artificial Intelligence http://www.racai.ro/wnbrowser/.
[25] Synset is the wordnet's basic unit: a set of synonyms which define a specific meaning, common to the members of the synset.
[26] www.ontologyportal.org

The first version of RoWN[27] was one of the outcomes of the BalkaNet European project[28] (2001–2004) at the Research Institute for Artificial Intelligence[29] of the Romanian Academy, in Bucharest, and the Faculty of Computer[30] of the ”Alexandru Ioan Cuza” University of Iași [Barbu-Mititelu, 2013].

Each synset contains a set of words (literals), each paired with a sense index, considered to be synonyms in certain contexts. Synsets (lexicalised concepts) can be considered to be nodes in a huge network, while edges are different semantic relations: hyponymy, hypernymy, meronymy, implication, cause, etc. (see Figure 4.1). Moreover, each RoWN synset is tagged by a domain label.

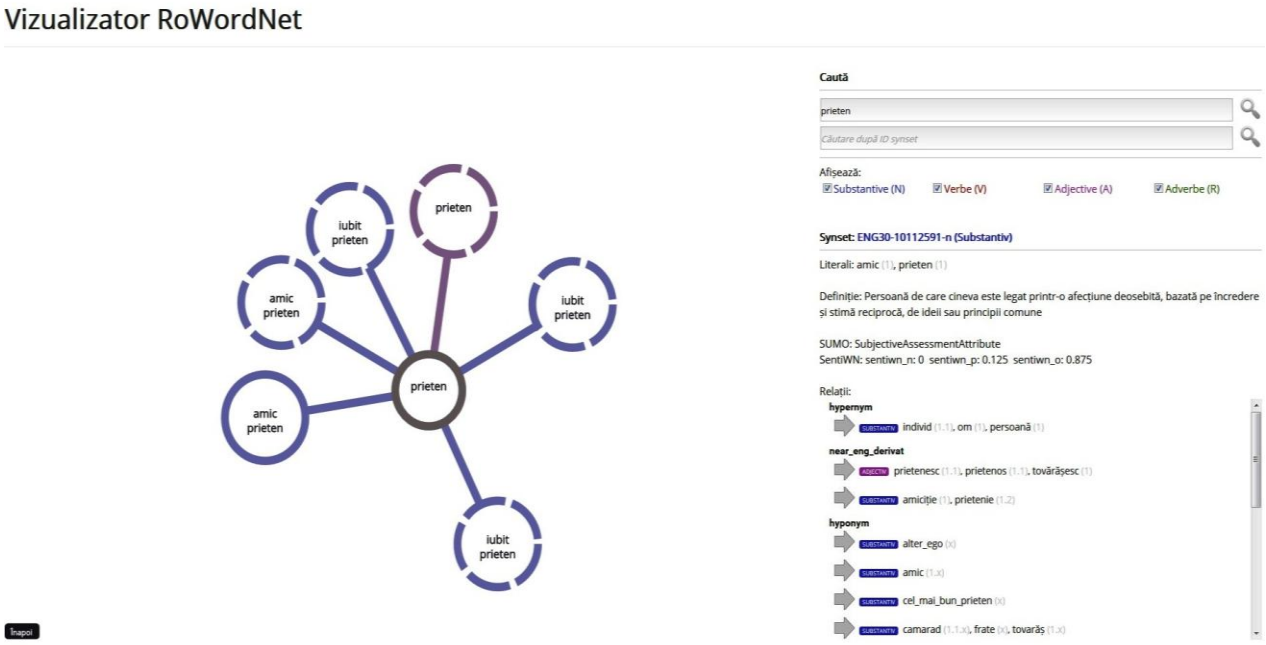


Figure 4.1 RoWordNet Vizualizer[31]

Three important resources have been used in the implementation, as described below:

- an *English-Romanian dictionary* – automatically extracted from a parallel corpus and then manually validated, containing 74.111 inputs;
- *EXPD* (*Explanatory Dictionary of the Romanian Language*), is a general dictionary of modern Romanian, containing about 56.000 entries[32]; it has been encoded in the

[27] Distribution of RoWN has been made through the project METANET4U.
[28] http://www.dblab.upatras.gr/balkanet/
[29] http://www.racai.ro
[30] http://www.uaic.ro/uaic/bin/view/Main/WebHom
[31] http://www.racai.ro/tools/text/rowordnet-visualizer/
[32] EXPD is a Romanian dictionary authored by the Linguistic Institute of the Romanian Academy. The number of entries applies for the 1996 edition of the dictionary. The last version (2002) has almost 100,000 entries.

XML format;

- a *Dictionary of synonyms* – the electronic version of the *Dictionarul de sinonime* [Seche & Seche, 1997], manually transposed in electronic form, encoded in XML and enriched with other series of synonyms automatically extracted from EXPD.

Two instruments were created especially for the development of RoWN, *WNBuilder* and *WNCorrect*[33] [Tufiș *et al.*, 2004].

I. *WNBuilder* - a graphical interface that allows the user (Figure 4.2):

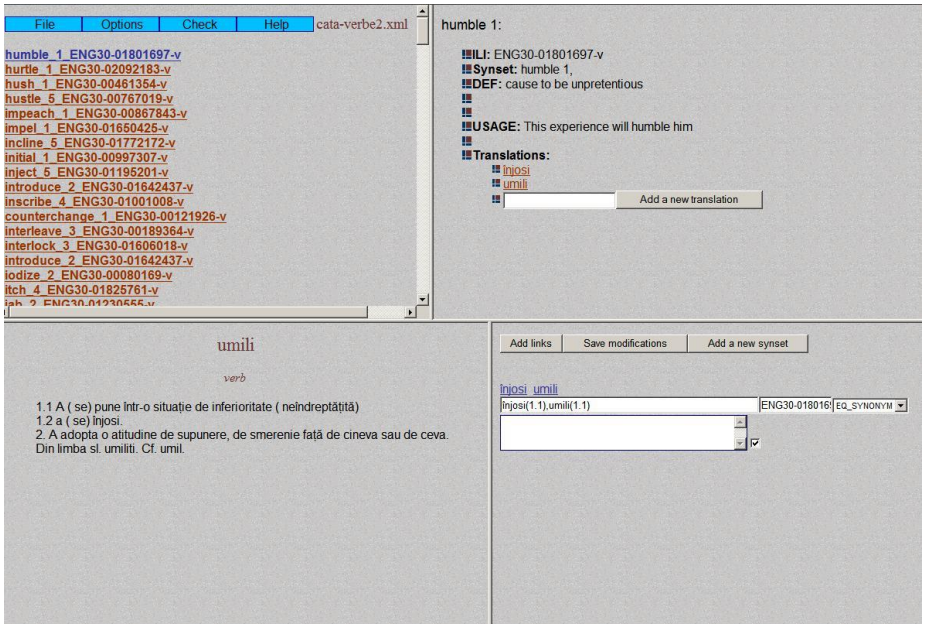


Figure 4.2 A working sesion with *WNBuilder*

- to access the available resources;
- to access the list of synsets from PWN for which synonyms have to be created in RoWN;
- to create synsets and associated glosses and to specify the sense number for every literal which is member of a synset;
- to importe the specific relations of each synset from PWN;
- to validate the syntax of the created synsets, by signaling the following types of errors: synsets without literals, literals without the sense number, repeated inclusion of a literal in the same synset, regardless of the sense number, and typos, as, for instance, lack of spaces, commas, etc.

An encoding scheme used in the *WNBuilder* system is exemplified below:

[33] Both WNBuilder and WNCorresct are implemented in Javascript and Perl runs under IE 6.0 or higher and are freely available on demand.

```
<ENTRY>
      <WORD>abandonat</WORD><POS>adjectiv</POS>
      <DEF>1. Care a fost părăsit.</DEF>
      <DEF>2. <USG>Despre copii nou – născuți</USG>Lepădat.</DEF>
      <ETYM>Vezi abandona</ETYM>
</ENTRY>
```

The <ETYM> tag is optional and it can be used to derive some lexical relations (here, there is a link to the verb from which the adjective is derived) [Tufiș & Barbu, 2004].

II. *WNCorrect* – a graphical interface that allows the user (Figure 4.3):

- access to EXPD;
- to access the synsets with conflicting literals, i.e. the same literal with the same sense number appearing in two or more synsets;
- to eliminate conflicts;
- validation of the synsets' syntax and their correction, in the same way as with *WNBuilder*.

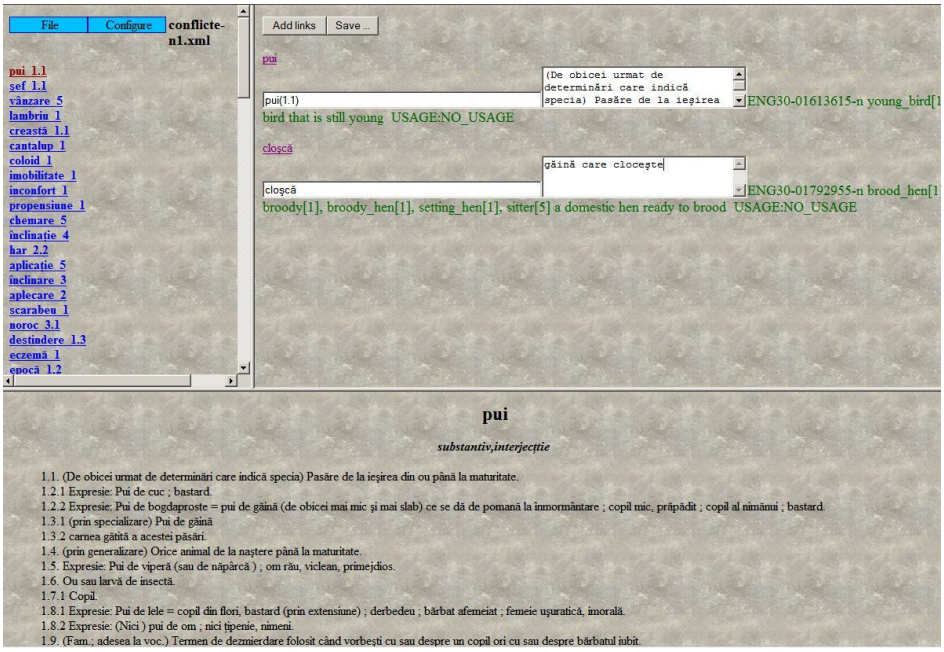


Figure 4.3 A working session with *WNCorrect*

RoWN is aligned with PWN, but has some differences:

- RoWN unlexicaled synsets: according to the thesis of specific semantics of languages [Coșeriu, 1966], some concepts are not lexicalized in some languages. If one concept that is lexicalized though one synset in PWN does not have a lexicalization in Romanian, the synset in RoWN has the form {nl(1)}. For example, the English noun *hugger*, with the meaning "the person who hugs", doesn't have an equivalent in Romanian, so in RoWN it has the synset {nl(1)};

- literals are randomly distributed in synsets because of the lack of a Romanian corpus annoted at the sense level, out of which one could extract data about the frequency of use of word senses;

- synsets in which appear identical literals (with different senses) are randomly distributed in the list of synsets of those literals (for the same reason as above);

- the numbering of senses of literals is done by copying the order of occurrance of the literal in EXPD. If the literal does not appear in EXPD, it receives as sense index *x*. If the literal appears in EXPD, but its sense that participate to the synset is not registered in EXPD, then it still receives as a sense number *x*; for example, the noun *clasă* (*class*) does not appear in EXPD with the meaning „elegance in dressing and behavior", thus *clasă* is assigned the sense number *x*;

Let's note a few uses of RoWN:

- *Semantic disambiguation.* [Ion & Tufiș, 2009; Ion & Ștefănescu, 2011] describe methods for word's sense disambiguation using wordnets. In the first mentioned paper, disambiguation is made in a multilingual context, using nets of aligned words; in the second paper, lexical chains connecting words in the sentence are used and the disambiguation is based on monolingual wordnets.

- *Automatic translations*. Existance of aligned-with-each-other wordnets allows the extraction of lists of translation. For example, bilingual dictionaries used in automatic translations. In this context, the multiword-terms of wordnets are very useful, which significantly improves the translation results.

- *Explanative dictionary*. Every synset has an associated gloss; commonly, in the Romanian wordnet this is reproduced from the Explanative Dictionary of the Romanian

Language in electronic format (EXPD). Thus, there where synsets have beed enriched with literals acquired from other sources than EXPD itsels, new entries can be added to the dictionary, copying the definition from the gloss.

- *Dictionary of synonyms*. Each synset is a possible input into a dictionary of synonyms. In an ordinary dictionary of synonyms, synonym words are listed without specifying the senses for which the synonymy relation is valid. Since wordnet operates the semantic distinction between literals, synonymy relations can be indicated for each of their semantic directions: if the synonymy between two words is valid for more of their senses, then these two words will appear together in a number of synonymy sets equal to the number of semantic directions they have in common.

- *Dictionary of antonyms*. In WordNet, there are two types of antonyms of adjectives: direct and indirect. Direct antonymy is established between literal senses displaying genuine opposition. Indirect antonymy is established by transitivity: if an adjective *a1* occurs in the same synset with an adjective *a2* and this one is in an antonymy relationship with an adjective *a3*, then an antonymy relationship will be set between *a1* and *a3* (similarly, if *a1* has an antonym *a2* and this one is synonym with *a3*, then an antonymy relationship will be set between *a1* and *a3*). In addition, applying these rules recursively, the antonymy relation can be extended even more.

- *Multilingual dictionary*. The existance of wordnets for different languages, all aligned with an InterLingual Index (ILI) and preserving similar hierarchies, invites for generation of multilingual dictionaries, in which, together with a source language term, equivalent terms in a number of other languages can be selected.

- *Syntactic dictionary*. The Romanian wordnet contains subcategorization frames for over 500 verbal synsets. A novelty in the formulation of these frames is the indication as constraints for each semantic role of a verb, of synsets (represented by pairs literal+sense), instead of mere literals.

Here is an example of a valence frame for the verb *to eat*:

(5) `{eat} nom*AG(person:1|animal:1)=acc*SUBSTANCE(food:1)`

The subject of this verb has the semantic role AGENT, and it can be satisfied by any member of the same synset as *animal* with sense 1, or by any member of the synset which includes *person* with sense 1, or any noun in the nominative case (nom) registered in wordnet as a hyponym of person:1 or of animal: 1. The direct object, in the accusative case (acc), has the semantic role SUBSTANCE and can be satisfied by the noun *food* with sense 1, or by any member of the same synset and any member of a hyponym synset in wordnet.

Semantic descriptions of this type, if extended over all +5000 verbs in WN, would be of a tremendous help in the realisation of a syntactico-semantic parser for the Romanian language. Such frames could be automatically extracted from wordnet and a parsed corpus of examples, provided a sufficient number of examples would be found for each verb (to generalise the frames adequately) and all words would be annotated for senses.

For verbs whose corpus instances are not sufficiently frequent or are absent, quotes from the thesaurus dictionary could be taken instead. In this case, however, the examples should also be filtered chronologically, in order to cope with evolution of language (syntax and semantic constraints could change in time).

4.3 The Thesaurus Dictionary of the Romanian Language (DTLR)

eDTLR – the Thesaurus Dictionary of the Romanian Language in electronic form[34] is a Romanian comprehensive lexical resource, heavely used in this PhD thesis (see Chapter 7). Every entry in the vocabulary of old forms, built in support of the research carried out for this PhD thesis, is searched for in eDTLR in order to improve their chronological data. We were interested in following the evolution of words over time.

eDTLR is an important lexical resource for the computerization of the Romanian language. It contains an impressive collection of textual sources, for the time being accessible only to specialists [Cristea *et al.*, 2007; Cristea *et al.*, 2011]. eDTLR is coded in the XML format. In fact, the users' interface (see Figure 4.4) reflects the complex architecture of the dictionary, transposed in the universal categories of TEI (*Text Encoding Initiative*)[35] standard to

[34] The project started in 2007 by a consortium structured from 7 partners: "Alexandru Ioan Cuza" University of Iași, (Faculty of Computer Science and Faculty of Letters), Romanian Academy (Institute of Linguistics "Iorgu Iordan - Al. Rosetti" Bucharest, Institute of Romanian Philology "A. Philippide" Iași, Institute of Linguistics and Literary History "Sextil Pușcariu" Cluj-Napoca, Research Institute for Artificial Intelligence Bucharest, and the Institute for Computer Science Iași).

[35] TEI is an encoding standard protocol for dictionaries building. More about encoding standards can be found in [Trandabăț & Gîfu, 2016]. See http://www.tei-c.org/index.xml, [15 May 2016].

represent the texts in digital format.

Figure 4.4 eDTLR graphical interface

eDTLR has a similar and compatible structure with online dictionaries for many different languages, as described below.

```
articol
     cuvânt titlu
     elemente de calificare
     sensuri   nr
          def
          cit
     informații privind pronunțarea
     informații morfologice
     variante grafice
     variante lexicale
     etimologie
```

Currently, only volumes corresponding to the new series of the Dictionary are accessible. The eDTLR database contains other volumes, including those from DA series (Academy Dictionary)[36], but they are in a processing process (parsing). After finalization of the processing of existing articles, these will be available online.

[36] DA (Dicționarul Academiei/the Academy Dictionary), in fact is Dicționarul Limbii Române/ the Romanian Language Dictionary was coordinated by Sextil Puşcariu between 1906-1944 (3071 pages, including the letters A-C, F-K, and a large part of L, D being under publication) [Seche, 1969]. It was based on the Phillipide's files and on the database of the Puşcariu and his team. [Puşcariu, 1913].

The eDTLR's entries structure is that of a tree, in which senses are recursively organised. We give below the sense structure of a verb: the sense tree contains 3 senses correspoding to level 3, one of them having a sub-sense correspoding to level 5 [Curteanu *et al.*, 2008].

```
<entry>
     <hw>VERB</hw>
     <senses>
          <marker level=”3”>1.
               <definition>…</definition>
               <marker level=”5”>◊
                    <definition>…</definition>
               </marker>
          </marker>
          <marker level=”3”>2.
               <definition>…</definition>
          </marker>
          <marker level=”3”>3.
               <definition>…</definition>
          </marker>
     </senses>
</entry>
```

Once completed, eDTLR [Cristea *et al.*, 2007] will be an important lexicographical resource for the Romanian language.

The dictionary meanings are included in phraseological units and the uses of the entry word are stylistically marked, but they aren't defined, only illustrated with quotes. These quotes can be used to form a collection of annotated examples. Examples are ordered chronologically, starting from the oldest, down to the most recent. This allows establishment of filiations, and tracking of diachronic changes suffered by the meaning of words.

4.4 Conclusions

In this chapter we have given a brief overview of the achievements and the lexical perspectives in language technology (in Romania) in general and of the language technologization in particular, which aims to highlight the importance of the lexical resources in NLP. Being such a vast field of investigation, language requires a coordination of highly qualified specialists in linguistics, artificial intelligence and cognitive science.

As it is expected, the unprecedented significance for the Romanian computational linguistics of the presented lexical resources (RoWN, eDTLR) significantly eases the efforts of different categories of users willing to find there more than simple information about the meanings of some words, therefore in need for complex language research: word sense disambuization and lexical choice, the automatic translation, diachronic research, etc.

Developing these two lexical resources is an ongoing endaviour. They can be used for a multitude of applications (including multilingual, as is the case with RoWN), sometimes in correlation. Both resources will be used in the applicative area (see Chapter 7) of this PhD thesis.

5. TEXT MINNING TECHNIQUES AND MODELS

This chapter includes some techniques and models belonging to the subdomain of Natural Language Processing (NLP) called ”Text Mining” (TM) [Gupta & Lehal, 2009], which we quickly revise, since we will make use of them in Chapter 7. TM is a knowledge-intensive process, which can be summarised as in Figure 5.1.

We have chosen only those methods that are directly related to our PhD research goal, namely to follow the language variation over time. The data in our case is a large collection of newspapers, chronologically ordered, starting with the second decade of the 19th century until today. This collection made first the subject of an epoch detection exercise, using as criteria major linguistic changes.

We briefly describe the text mining specific problems and general aspects in the supervised, unsupervised, and semi-supervised learning with implication for word/language similarity.

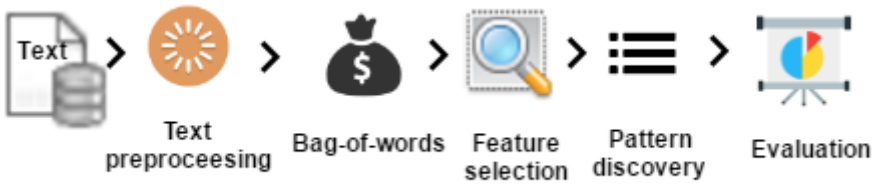


Figure 5.1 Text mining process

TM techniques and models have been largely used in NLP to discover important language characteristics that define the epochs of the documents in order to extract relevant information about them [Gupta & Lehal, 2009]. But the other way round is also a plausible hypothesis, because in order to apply TM methods, NLP techniques are being used to extract features. To prepare textual documents for TM analysis, they usually pass through various processing phases, such as: eliminating insignificant words („stop words”), finding word roots or lemmas, identifying parts of speeches, recognising named-entities, etc. Moreover, time series analysis is becoming more and more widely used [Kao & Poteet, 2004; Mihalcea & Năstase,

2012; Popescu & Straparava, 2013/2014; Ciobanu *et al.*, 2013], etc.

5.1 Supervised, unsupervised and semi-supervised learning

Machine learning refers to the way learning techniques are used, consisting in unsupervised learning (clustering or neural networks), supervised learning (classification) and semi-supervised learning, the last one being an extension of the concept of supervised learning.

A machine learning process is sketched in Figure 5.2. As can be seen, the learning mechanism is strictly dependent of the Data component, namely – the relation between the training data and the development data (e.g., a language model).

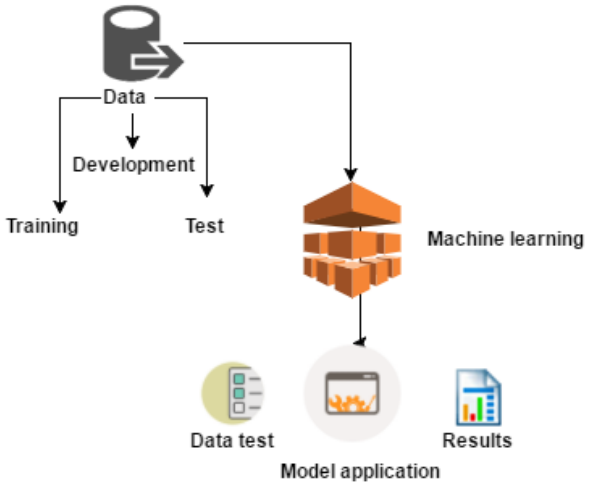


Figure 5.2 Machine learning specificity

For ***unsupervised learning,*** (we will refer below to Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)), there is no need for training data. The human effort is zero, which will be reflected on the data accuracy, but greatly increases the complexity of the algorithm, supposed to be general and to necessitate higher response time. The goal of unsupervised learning is to discover some intrinsic structures in the given data.

For instance, when we talk about semantic similarity, we refer, theoretically, to a metric that reflects the matching of two strings of text, expressed by a number of features. In NLP are well known document (or text) clustering [Li & Chung, 2005] and information retrieval systems [Grossman & Frieder, 2004].

For ***supervised learning***, training data is large (here below we will refer to NaïveBayes, Maximum Entropy, Support Vector Machine (SVM) and Long Short-Term Memory (LSTM)).

The human effort to build accurate data collections is very high, somehow counter-balanced by a decrease in the complexity of the algorithm. Supervised learning or clasification or inductive learning in machine learning has bigger success in NLP applications, being similar to human learning: considering our past experiences (or the acquired knowledge) and our skills to observe the reality (or cluster the acquired examples) we are able to improve our ability to face new challenges. Because the computers have no "experience", the supervised learners learn from training data, which represent the past experience in some real world applications.

For ***semi-supervised learning***, on the basis of some training data, others are produced automatically. Here, the human effort is moderate, the accuracy being related to the complexity of the algorithm [Chapelle *et al.*, 2006][37]. If the algorithm is very complex, for a small training data, the accuracy is supposed to be high. The semi-supervised learning is somewhere on the border of supervised learning (uses only labelled data sets) and unsupervised (no data driven labelling).

5.2 Text classifiers

Traditionally, text classifiers require labelled training examples. Unfortunately, labelling is a time consuming process, because it is done manually in a large proportion, usually by human experts. This is the reason why the community of NLP researchers investigates various ways of semi-supervised learning [Blitzer & Zhu, 2008] to reduce the manual labelling process: self-training methods [Blum & Mitchell, 1998; Haghighi & Klein, 2006], graph regularization methods [Zhu *et al.*, 2003], structural learning [Ando & Zhang, 2005], etc.

Text classifiers are machine learning techniques for text documents classification based on an agglomerative algorithm, necessary to frame a text document in a predetermined number of categories. [Rebedea *et al.*, 2008].

In this section, we will focused on two popular classifiers in NLP, NaïveBayes and Maximum Entropy [Lewis, 1998; Nigam *et al.*, 1999] which have been proven to be efficient in diachrony (e.g. old text documents classification in certain epochs). These models operate with texts represented as word counts, implying linear decision rules expressed in log-space in order to measure the relative degree of association between a class-word pair and the independent

[37] The semi-supervised learning algorithms use at least one of the following assumptions: (1) *cluster* – if a few points are in the same cluster, probably they belong to the same class; (2) *smoothness* – if in a high-density region, two points are close (belonging to the same cluster), then the corresponding outputs should be close; (3) *manifold* – in a high-dimensional space, the data lie on a low-dimensional manifold.

parameter allocated to it. [Juan *et al.*, 2007]. MaxEnt technique is closely related to a NaïveBayes which aims training process, how features weights are set, and dependecny on features. [Klein & Manning, 2003].

5.2.1 Naïve Bayes – a generative model

The Bayesian decision theory was mentionned for the first time in Pattern Classification and Scene Analysis [Duda & Hart, 1973]. Indisputable, the Naïve Bayes[38] binary classifier (implemented in the WEKA[39] (Waikato Environment for Knowledge Analysis) library is one of the most popular and simplest statistical models based on trained data. It is a generative classifier that uses a very simple representation of the document (called "bag of words").

This method of Bayesian learning (Figure 5.3) is applied for learning tasks (e.g. to predict the period of which a text document belongs) where each instance x (e.g. a text document) is described by a conjunction of features, f (Formula 5.1).

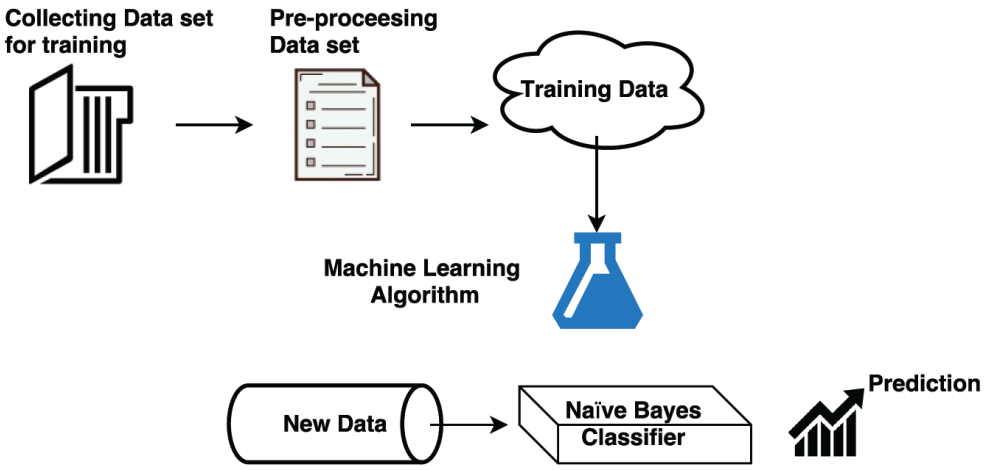


Figure 5.3 A diagram of a model building procedure for pattern classification

It starts from a training data set that requires labelled data (being a supervised learning) where each instance x is described by a set of features (f_1, f_2, . . . , f_n).

$$x = \sum (f_1, f_2, \ldots, f_n) \qquad (5.1)$$

[38] Naïve Bayes is a simple ("naïve") classification method based on Bayes rule.
[39] http://www.cs.waikato.ac.nz/~ml/weka/

The system is requested to make a prediction (p) about the target features, which is the class to which the new instance belongs. The Bayesian approach for classifying the new instance consists in assigning the instance (x_i) to the most probable class (c_i) having a set of features.

$$x_1 \rightarrow c_1$$
$$x_2 \rightarrow c_2$$
$$\cdots$$
$$x_n \rightarrow c_n$$

Each class is checked for each new instance, giving a probability, so that the sum of probabilities to be 1. (Bayes formula 5.2). Note, the way they have defined these probabilities builds a language model.

$$\sum p(c_i|x) = 1 \qquad (5.2)$$

Naïve Bayes invented the calculation of the *conditional probability* or the *posterior probability*, based on Bayes' formula to which added conditions, using `if`. If we have no conditions, the probability is called *static* or *a priori*.

The efficiency of this model has been proved in various situations: spam filtering [Goodman *et al.*, 2005; Bratko *et al.*, 2006], information retrieval [Robertson & Sparck Jones, 1976], sentiment analysis [Narayanan *et al.*, 2013], language identification [Winkelmole & Mascardi, 2011], etc.

The Naïve Bayes model implies a normative framework for representation and motivation on the issues of decision making under conditions of uncertainty. It proves to be a very attractive choice when it comes to a large set of independent variables. Although it is fast and easy to implement, most popular statistical software packages do not include it. Here are the reasons:

(1) the probability classes are biased, which could be a real problem for modelling applications where the focus falls not necessarily on the classification task;

(2) the relations between the dependent variables and predictions are estimated in an isolate mode, without taking into account the covariance between predictions.

As a consequence, the Naïve Bayes classifier is not able to approximate the multivariate regression functions and, as a tool for data exploration, it does not add any more information than an univaried analysis.

In specific fields, the performances of the Naïve Bayes classifier were comparable with those of neural networks or of decision trees learning methods. As shown in [Domingos & Pazzani, 1997], in the presence of dependence characteristics, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class.

In this PhD thesis, the Naïve Bayes method is used for obsolete word recognition, labelled with the `NotInDict` tag, depending on two parameters, the geographic region and the epoch, statistically detected (more in Chapter 7) in the curent corpus (Chapter 6). Many scientists (see for instance [Rennie *et al.*, 2003]) invoked the performance of this classifier on text classification tasks. We have used it in recognizing the two classes for new texts.

5.2.2 Maximum entropy modeling

In this section, we will present the Maximum Entropy text classifier, known as the MaxEnt classifier, a commonly used classifier for Natural Language Processing tasks, first introduced by Berger *et al.* [1996] and Della Pietra *et al.* [1997] in statistical estimation and pattern recognition. At the same time Ratnaparkhi [Ratnaparkhi, 1996; 1997] understands the importance of this model in POS tagging, combining various contextual elements. Moreover, the MaxEnt model performed well in text categorization [Nigam *et al.*, 1999], parsing [Charniak, 2000], etc. Noteworthy is that the MaxEnt classifier has good results even when the training corpus is small.

MaxEnt is a discriminative or conditional model, meaning that it learns the conditional probability distribution `p(y|x)`, which should be read "`the probability of y given x`".

To solve the textual entailment, Malakasiotis [2010] describes a machine learning approach that uses a MaxEnt classifier. The system is based on a series of string similarity measures as features, in order to recognize the textual entailment.

Ren and his colleagues [Ren *et al.*, 2010] also proposed a system of classification based on MaxEnt model to recognize textual entailment, taken into account: the lexical (Bag-of-Words, Jaccard similarity, Longest common substring, Levenshtein distance), syntactic (Bag-of-Dependencies, Descendent Relation Features, Combined Verb Descendent Relations, Combined Subject Descendent Relations, Combined Subject-to-Verb Relations and Object Relations) and semantic features (using WordNet) and some specific features (relations detection

between named entities using Wikipedia, idea inspired from [Iftene & Balahur-Dobrescu, 2008].

MaxEnt has been successfully applied to diachrony, where it is put at the base of a machine learning algorithm (see Chapter 7). The MaxEnt classifier is useful for automatically correlating different forms of a word (e.g. to estimate the best word pairs), being suitable for bootstrapping (a sampling with the replacement included in test data), in order to increase the quantity and quality of the training data.

By using the contemporary language as a pivot in [Gîfu & Simionescu, 2016; Gîfu, 2016a] the data (e.g. old words that belong to versions of one language) is analyzed and compared from various epochs, using MaxEnt model in a supervised learning, from the perspective of transformation rules. By maintaining an inventory of these rules, automatic correspondence between the words of different lexicons can be traced.

5.3 Artificial neural networks

It is known that the people's thinking mode is based on neural connections. In other words, our thinking does not start from zero, every word that we have thought or spoken is based on the understanding the preceding words. Since 1943, when Warren McCulloch and Walter Pitts built a model which simulate the biological neurons based on some resistors and amplifiers [Mc Culloch & Pitts, 1943], the interest for artificial neural networks (ANNs) has visibly increased.

The main advantage of ANNs is the structural autonomy. They can provide an accurate estimation of the dynamics of complex systems, without any information about the systems' components. Summarizing, ANNs consists in: parallel processing, inductive reasoning, dynamic and internal learning capacity based on the knowledge accumulated through examples, using specific learning algorithms, the approximate correlation of the data for retrieval of the conclusion earlier learned, tolerance to errors or inaccurate data, etc.

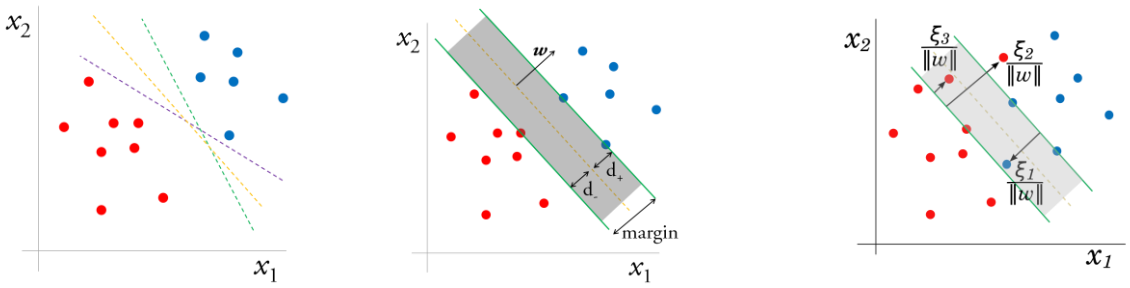
We present below two neural networks, SVM (Support Vector Machine) and Long Short-Term Memory (LSTM) that we used for language similarity, whose results have demonstrated the effectiveness in the classification of texts. These probabilistic models are based on the same principle described by Salton [Salton, 1971]. Each document represented as vectors of words is optained via weighted bag-of-words method.

5.3.1 Support Vector Machines (SVMs)

An approach based on space density computations similar with SVMs was introduced in 1975 by Salton, Wong and Yang [Salton *et al.*, 1975]. They were interested to choose an optimum indexing vocabulary for a collection of documents. First mention about SVMs was in 1979 by Vapnik and colleagues. Two years later, were published the first works by a group of researchers led by Vapnik [Vapnik, 1995; Vapnik & Cortes, 1995], who insist entirely on the pattern recognition problem (e.g. handwritten digit recognition). A typical application of pattern recognition problem is text categorization [Joachims, 1997; 1999; Burges, 1998], where it is demonstrated that the SVM classifier obtains better results than other classifiers. Furthermore, the uniqueness of SVM consists, in fact, that they do not incorporate the problem-domain knowledge.

Conceptually, SVM "is a linear machine, equipped with special features, and based on the structural risk minimization (SRM) method and the statistical learning theory" [Gorunescu, 2011:234]. Known as one of the most popular models of classification, the basic problem is to identify a hyperplane that ensures the separation of classes (see Figure 5.4). [Coates *et al.*, 2011]. The training process presumes the estimation of parameters of hyperplanes (and possibly of some core functions) used in the transformation of a nonlinear separation problem into a linear separation one. The classification consists in calculating the values associated with the separator function(s).

Often, the question in SVM is: which separator hyperplane is better? The answer should be searched in the one that provides a good generalization capability. As will be seen in Chapter 7, we will use this method to recognize the linguistic characteristics of an epoch.



a) Maximazing the margin b) Formulating the margin with geometry c) Optimazing problem

Figure 5.4 A binary classification problem (reproduced from [Yeh, 2015[40]])

[40] http://efavdb.com/svm-classification/ - online course

The models based on SVM come with the idea of representing a document through support vectors without to restrict the number of features. Too many features imply imprecise results. It solves the problem of dimensionality without affecting the efficiency or restricting the universality.

In principle, each text becomes a vector in an *n*-dimensional space. To each word belonging to the collection of texts corresponds a size of the vector, the numerical values corresponding to each document for each word being determined using a weighting scheme (that considers the word's presence or absence, its frequency in text, and the TF-IDF – the terms frequency multiplied by the inverse of the document frequency where the terms appear).

Note that SVM recently began to be used in diachrony. We remind a study of syntactic productivity in diachrony [Perek, 2014] showing that it is mainly determined by semantics. This work uses the bag-of-words approach based on word co-occurrence frequencies within a given context window. In principle, it measures the semantic similarity between verbs, within a corpus of 464 million words of American English, called COCA. The productivity depends on how the existing semantic space refers to the new lexical term and if this relation is attested by several lexical terms.

One of the SVM studies that will underpin the language similarity module of this PhD thesis belongs to Popescu & Strapparava [2015], which describe the diachronic text evaluation task. They predicted texts epochs when important linguistic differences were found in the written press of the period from 1700 till 2010. Then, they have used a number of statistical methods for dating a text as a stylistic classification problem. For this, they have trained a multi-class SVM model, using a set of stylistic features (character frequency, POS-tag, word, n-grams, and syntactic phrase-structure rule occurrences) extracted from their corpus.

5.3.2 Long Short-Term Memory (LSTM)

In the middle of 90s, two German researchers, Hochreiter & Schmidhuber [1995] proposed a special kind of recurrent neural network (RNN), called Long Short-Term Memory or LSTM, a set of recurrently connected subnets. It was designed to avoid the long-term dependency problem, being capable of recognizing both long and short patterns in sequences of data.

In the LSTM training architecture is prevented the *exploding gradients* and the *vanishing gradients* problems introduced and described in [Bengio *et al.*, 1994; Hochreiter *et al.*,

2001], both known for efficiently training RNN. The extension of gradients is influenced by their weights and derivates. If one of those two factors is under 1, we speak about *vanishing gradients*. If those factors are above of 1, then it could happen the *exploding gradients*.

Below, the diagram depicts a recurrent network in an ongoing recognition process. The represented units are filled with different colors based on how sensitive they are to the entrance at time 1 (black - high and white - low). The figure 5.5 shows that the intensity of the input decays exponentially in time.

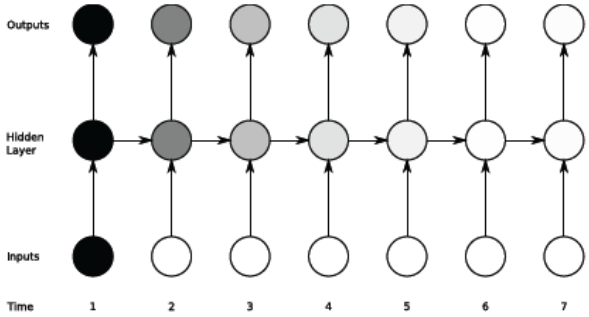


Figure 5.5 Illustration of the vanishing gradient problem (reproduced from [Graves, 2008:33[41]])

The solution is given by training RNN with gradient-based methods and backpropagation, used in training (supervised learning). Because to this backpropagation through time, the LSTM models perform in practice better than other RNNs. By using memory cells to store information, an LSTM memory block is efficient at training to remember the values for long time series (Figure 5.6).

LSTM networks are composed of multiple LSTM cells connected to each other. Each cell has the following architecture: to its base is a linear unit, or a neuron, that in every moment sums up all the information that it receives from other neurons with which it is connected. To retain an information for a longer time, it is necessary to implement a circuit to simulate a memory cell. Thus, the basic cell is linked to other cells: *forget gate* (when the value is 0, the basic cell maintains its status), *write gate* or *input gate* (when the value is 1, the rest of the network can change the value of the basic cell), and *read gate* or *output gate* (when the value is 1, the rest of the network can access the information stored in the basic cell). Thus, the cell has a recurrent connection with fixed weight 1.0. All three mentioned gates can be trained to learn. Moreover, they store information in memory, decide how long and when to read from the rest of the network and controls the cell through multiplicative units (small circles). The input (write) and

[41] http://deeplearning.cs.cmu.edu/pdfs/1104/Supervised_Sequence_Labeling.pdf

output (read) gates check the input and output of the cell. For indication of the places, are applied the non-linear cell squashing functions *g* and *h*.

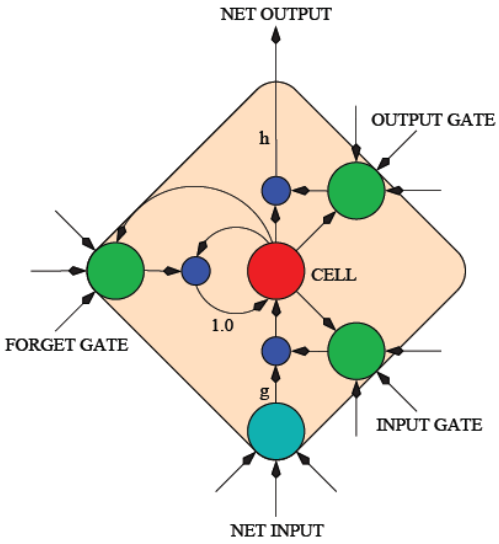


Figure 5.6 Illustration of an LSTM memory block with a single cell (reproduced from [Graves, 2008:34[42]])

The model has succeeded to solve a part of the fundamental problems of traditional recurrent neural networks (RNNs) on tasks involving long delays between significant input events [Hochreiter & Schmidhuber, 1996; 1997a; b], including in the following application areas: time series prediction [Mozer, 1989], speech recognition [Dahl *et al.*, 2012], handwriting recognition [Graves & Schmidhuber, 2009; Frinken *et al.*, 2012], etc. Those tasks, however, did not require the network to extract information conveyed by the duration of intervals between events.

Increasingly more applications are based on LSTM models, very promising for the field of language modelling, first mentioned in [Sundermeyer *et al.*, 2012]. For instance, Mikolov and his colleagues [Mikolov *et al.*, 2015] propose a model able to predict the next token in the sequence given its past. They show that LSTM models allows for significantly more parameters for the same size of hidden layer, making the comparison with other models reckless.

Our interest in using the LSTM model is connected to language identification. In this sense, Gonzalez-Dominguez and his colleagues [2014] show that LSTM can be useful in modeling temporal dependencies in the acoustic data, being helpful in learning relevant long-range discriminative features for language identification. Another interesting paper presents an

[42] Idem.

open-source LSTM system for automatic language recognition of short utterances in 8 target languages [Zazo *et al.*, 2016]. They adapt the previously mentioned approach to explore different aspects that affects LSTM performance, implementing and extending the insights of that work, using open source software and much more limited resources.

In language similarity, LSTM proves also its efficacy. Thus, a more recent study shows how LSTM model can perform remarkably to evaluate the semantic similarity between sentences using labelled data [Mueller & Thyagarajan, 2016], because of its hidden units to encode different features of each sentence. Authors have used word embedding vectors, including synonymic information. This interest for text understanding by modelling the semantic similarity between sentences was preceded by another [Marelli *et al.*, 2014]. However, a problem is still unsolved due the sentences length diversity, the complexity of their structure, the bag-of-words /TF-IDF approaches. In [Mihalcea *et al.*, 2006] is emphasized the fact that, in the measuring of the semantic similarity, bag-of-words model ignores important relationships in the sentence structure (e.g. the role played by different arguments in a sentence context or dependencies of words). As an alternative, the standard LSTM architecture can be generalized to tree-structured network topologies for representing sentence meaning over a sequential LSTM. [Tai *et al.*, 2015].

To our knowledge, so far, the LSTM model has not been used in diachrony. This PhD thesis advances the first study in this respect, in order to find the best model in the contrastive analysis between cognate languages (see Chapter 7).

5.4 Probabilistic topic models

The probabilistic topic models are based on the principle that a collection of documents (e.g. printing press articles) include a mixture of randomized topics where every one represents a statistical distribution over words. Such models have an extremely important role in various NLP applications. [Hofmann, 1999; 2001; Griffiths & Steyvers, 2002; 2003; 2004; Blei *et al*. 2003].

We present below two probabilistic models that can be used in our contrastive analysis in order to highlight the similarities between different corpora from a semantic perspective: the Latent Semantic Analysis (LSA) and the Latent Dirichlet Allocation (LDA). They are based on the bag-of-words assumption [Manning & Schütze, 1999]. Several experiments until now have proved that they can be used to complement each other, in the sense that the semantic relations are more susceptible to be detected when two approaches are combined, after a

normalization [Dascălu *et al.*, 2013a].

In this sense, we will show that the two models provide together a complex picture of the language similarity, by calculating a number of textual complexity indices that reflect the differences in writing style, ranging from sentence and paragraph structure to word entropy and cohesion [Gîfu *et al.*, 2016].

5.4.1 LDA – a probabilistic topic model

LDA (*Latent Dirichlet Analysis*) is a generative probabilistic model based on topic distributions [Blei *et al.*, 2003; Blei & Lafferty, 2009].

LDA considers that the documents are a collection of words and every word from a document is a sample of a mixture of models. Each component of this mixture can be seen as a *topic*. As a consequence, every word is generated by a single topic. If a document is a collection of words, these words could be caused by various topics. Each document is represented as a list of different proportions for each component of the mixture, so it is reduced to a probability distribution on a fixed set of topics. LDA offers an inference mechanism of a topics structure (classes of concepts) through a probabilistic generative process.

A topic is modelled by a Dirichlet distribution [Kotz *et al.*, 2000] over the vocabulary (the space of a multivariate distribution of words from the training corpora) in which thematically related concepts have similar occurrence probabilities. According with Mușat *et al.* [2010] LDA is based on a hierarchical and generative process. The hyperparameters α and β are the foundations of two Dirichlet distributions. The first Dirichlet distribution controls the mixtures model for each document in the collection *D*. The second Dirichlet distribution controls the generation of the words' mixture for each of the *K* topics. So, each topic is a distribution of the W words from the vocabulary.

The generative process is the following: for each word $w_{d,i}$, where d indexes documents and i – words, from our vocabulary, it generates a topic z dependent on the mix θ assigned to the document d and then it generates a word from the topic z. It is worth noting that those words that have no special relevance (articles, prepositions) will have approximatively the same probability for each class or can be placed in a special class. Determining parameters θ and z, and sometimes the hyperparameters α and β, could be difficult, because the *conditional probability* cannot be calculated precisely. This is why they are using different approximation

algorithms, such as Markov processes, etc. The specificity of the LDA model, as indicated in [Rivasseau, 2003], consists in associating a Markov chain with some topics or clusters.

A probabilistic approach like this one presents a number of advantages and disadvantages. The theoretical foundation is well known in Bayesian statistics and has been widely used. It is designed to create inferences on new documents.

LDA does not show as output a center for each cluster, but a distribution of the documents according to topics. This fact presents a number of difficulties in associating a name legibly. Recent works have been trying to find methods to associate topic names found by LDA using *n*-grams [Wang *et al.*, 2007].

Like the other models presented here, this probabilistic method does not solve the classical issue of finding the global optimum and the number of topics K. For the second issue, some methods have been proposed using statistical models presented in [Rodriguez, 2005]. The author describes three criteria for model selection in order to improve the optimization of a global topological quantity.

The algorithm LDA opens the way to solve various complaints related to the original model: topic trees extracted (hLDA) [Blei *et al.*, 2004], inducing a correlation structure between topics [Lafferty & Blei, 2005], modelling dynamic topics in time [Blei & Lafferty, 2006], using *n*-grams instead of simple words to describe topics [Wang *et al.*, 2007], etc.

A few applications based on LDA include the following tasks:

(1) Question answering (Q/A). We remind here a paper that introduces new similarity measures to discover deeper relationship between q/a pairs based on the LDA model [Celikyilmaz *et al.*, 2010]. The argumentation consists in the fact that the similarities can be characterized better if the semantic similarity measure is defined by topics on top of lexico-syntactic features.

(2). Sentiment analysis. Many works use the popular LA in the sentiment classification issue. For example, a paper by Lu *et al.* [2013] proposes a semi-supervised LDA method. That means that both labelled and unlabelled data are used in training the LDA model, fact that reduces the cost of manual labelling. Furthermore, the training process is more flexible when two ways of labelling are used, which brings an advantage over other methods in terms of multi-label classification.

(3). Cross-language text similarity. Yuan & Qian [2015] used the LDA model in cross-language text similarity computation to obtain Tibetan-Chinese comparable corpora. Although LDA is useful, some factors influence the correction of cross-language text matching (the accuracy of Tibetan word segmentation, incomplete Tibetan-Chinese dictionary, the particle size

of Tibetan and Chinese word segmentation).

5.4.2 LSA – a vector space model

To measure the semantic similarity between sentences or paragraphs we have used the Latent Semantic Analysis (LSA) model. It is based on a vector-space representation of semantics in order to depict the co-occurrence relations between terms and documents [Deerwester *et al.*, 1990; Berry *et al.*, 1995; Landauer & Dumais, 1997]. Same as LDA, LSA uses a bag-of-words approach, thus inheriting the limitations encumbered by ignoring the order of words.

Moreover, the LSA model uses a mathematical technique called Singular Value Decomposition (SVD) to create a new abstract vector-space by rewriting a term-document matrix. The LSA vectors are generated after the vectors projection obtained through SVD [Golub & Kahan 1965; Golub & Reinsch 1970; Landauer *et al.*, 1998a] of the TF-IDF matrix, used for evaluating similarity between terms and documents, now indirectly linked through concepts [Landauer *et al.*, 1998b; Manning & Schütze 1999]. The resulting vectors can be used to determine the proximity of terms by applying a cosines type similarity [Landauer *et al.*, 1998a] or other vector-algebraic metrics.

Based on the previous descriptions, the *ReaderBench* multilanguage platform for French and English operates [Dascălu *et al.*, 2013b]. It incorporates the English WordNet lexical database [Fellbaum, 1998; Miller 1995, 1998, 2010] and his French version, WOLF[43] [Sagot, 2008; Sagot & Darja, 2008]. Since 2015, this platform was adapted for the Romanian language (one of the contributions of this PhD research) [Dascălu & Gîfu, 2015]. Practically, it permits the evaluation of three main text features: the cohesion between various text fragments, the identification of strategies for reading and the evaluation of textual complexity.

LSA assumes that similar and/or related words in terms of meaning will occur in similar text contexts [Ngoc *et al.*, 2014]. The main problem, when we want to build a LSA matrix, is the large volume of data, which are analyzed in an unsupervised manner. The rows of the matrix are single words, and the columns are paragraphs or documents. The content of the matrix corresponds to the number of words from a paragraph / document.

In addition, LSA cannot capture polysemy: "Each occurrence of a word is treated as having the same meaning due to the word being represented as a single point in space" [Rhaman,

[43] WOLF - WordNet Libre du Français, http://alpage.inria.fr/~Sagot/wolf.html

2010:15]. This clearly introduces a very severe limitation, being clear that a word could have one or more senses and its current use depends on the context.

Example: *Ea o **presa** pe mamă să plece din casă. Azi a venit **presa** mai repede*. (En. *She **presses** on her mother to leave the house. Today the **press** arrived earlier*.)

For our diachronic survey, LSA model has proven to be efficient only for documents (newspaper articles) after 1941, due to important differences in word structure (old, obsolete forms not present in Romanian lexical resources used in the preprocessing phases) (more in Chapter 7).

5.5 Conclusions

In this chapter, we have briefly presented the characteristics, advantages and disadvantages of the most known text mining techniques and models frequently used in various NLP applications, with a special emphasis on diachrony. In our case, text mining seeks to extract useful information from a corpus (journals collections monitored over time) through the identification and exploration of interesting patterns.

In order to better understand language similarity from the four different corpora (a collection of publications from Moldavia, Wallachia, Transylvania and Bessarabia), we combined the techniques described above, emphasizing especially the words repetition and the normalized number of occurrences, with semantic distances extracted from ontologies or based on semantic models as LSA, LDA, etc.

We believe that the statistical modelling of texts is a generous research direction in full ascension. Furthermore, all these probabilistic systems require new methods of evaluation and interpretation of the results. Why? Because the results are often affected by noise, special properties of the words (such as polysemy), the dimension of a corpus, etc.

6. THE ROMANIAN LANGUAGE JOURNALISTIC CONTEXT – 1829-2015

In this chapter we briefly present the journalistic writing style from Romania and Bessarabia, on which we have built the corpus used in this PhD thesis, underlining also their main peculiarities. This context may be noticed based on a chronological criterion, which the literature records as the main indicator in the analysis of the evolution of the written press [Roquette, 2007; Petcu, 2009]. The corpus answers the conditions enounced in [Cristea & Tufiș, 2002], being a structured collection of texts in the Romanian language, in electronic form, that covers the journalistic style, published in four geographical areas (the Romanian provinces of Transylvania, Moldavia and Wallachia, and the province of Bessarabia), covering a time period extending from 1829 till 2015, and themes from culture, politics, religion, miscellanea, etc.

6.1 The phases of Romanian written press between 1829-2015

The present section proposes a brief description of the Romanian written press history taking into consideration some phases identified by theoreticians based on a series of spatial-temporal and contextual-political descriptors (see, for instance, [Diaconescu, 1974; Hangiu, 2004; Ilincan, 2010]).

The Romanian press of the first half of the 19th century contains pages from the most diverse areas: history, politics, literature, religion, etc. As the journalistic phenomenon was developing, representative publications appear in the second half of the 19th century (*Convorbiri literare* – Moldavia, *România liberă* – Wallachia, *Telegraful român* – Transylvania), whose profile shape according to the preferences of the editors directly involved and to subscribers' options.

6.1.1 A brief history of Romanian newspapers included in the corpus

The Romanian magazines (newspapers, journals) played an important role in the process of

literary Romanian language modernization, especially in the 19th century, exemplified and analysed, for example, in Andriescu's studies [1977; 1979], which we used in the explanation of the deflections of the old language, as identified in the corpus (see Table 6.1).

For exemplification, we briefly describe the main press events (*e.g.*, the issues of some publications) till the beginning of the 20th century, summarising a historical journey of the Romanian press, very well treated in the mentioned studies.

Table 6.1 The collection of the Romanian newspapers from PhD corpus

Publication	**Region**	**Year(s)**
Albina românească	Moldavia	1829, 1832, 1844, 1869
Convorbiri literare	Moldavia	1868, 1870, 1893, 1897, 1989-1991
Curierul. Foaia intereselor generale	Moldavia	1880
Constituționalul	Moldavia	1885
Basarabia reînnoită	Moldavia	1907
Lumea. Politică și Socială	Moldavia	1919
Lumea nouă. Politică și Socială	Moldavia	1919
Libertatea. Organ politic muncitoresc	Moldavia	1933
Contabilul	Moldavia	1934
Chemarea	Moldavia	1940
Chemarea Moldovei	Moldavia	1990
Bună ziua, Iași	Moldavia	1999, 2013
Ziarul de Vrancea	Moldavia	2009
Monitorul de Vaslui	Moldavia	2011
Evenimentul regional al Moldovei	Moldavia	2012
Imparțial	Moldavia	2014
Ziarul de Vrancea	Moldavia	2015
Curierul românesc	Wallachia	1829, 1875
Pressa	Wallachia	1877-1878, 1880
România liberă	Wallachia	1877-1878, 1880
Romanulu	Wallachia	1877-1878, 1880
Timpul	Wallachia	1877-1878, 1880, 1883
Literatorul	Wallachia	1880
Albina	Wallachia	1883-1884
Desteptarea. Foaie pentru popor	Wallachia	1911
Adeverul	Wallachia	1919, 1925, 1933, 1935
Curierul artelor	Wallachia	1919
Dimineata	Wallachia	1919, 1928
Universul	Wallachia	1919-1920, 1925, 1929, 1933-1934, 1936, 1938
Viitorul	Wallachia	1926
Curentul	Wallachia	1928-1929, 1932-1933, 1935, 1937-1940
Universul literar	Wallachia	1930
Dimineata	Wallachia	1931
Adevarul literar si artistic	Wallachia	1937
Scânteia	Wallachia	1947, 1989

România literară	Wallachia	1989-1991
Dimineața copiilor	Wallachia	1993
Evenimentul zilei	Wallachia	2014, 2015
Gândul	Wallachia	2014
Ziua	Wallachia	2014
Ziua news	Wallachia	2014
Ziua veche	Wallachia	2014
Organulu Luminarei	Transylvania	1847-1848
Gazeta de Transilvania	Transylvania	1852, 2009-2010
Telegrafulu Romanu	Transylvania	1853, 1865, 1873, 1875, 1877, 1883, 1888-1890, 1895, 1919
Foaia pentru Minte Anima si Literatura	Transylvania	1854, 1860, 1865
Federatiunea	Transylvania	1856, 1868, 1869, 1870
Transilvania	Transylvania	1866, 1869, 1876, 1887
Gura Satului	Transylvania	1869-1870, 1880-1881
Albina	Transylvania	1870
Familia	Transylvania	1881, 1890, 1894, 1935
Aradu	Transylvania	1890, 1930-1931
Patria	Transylvania	1919
Chemarea tinerimei romane	Transylvania	1926
Dreptatea	Transylvania	1926, 1929
Curierul crestin	Transylvania	1933-1934
Noutatea	Transylvania	1933-1934
Vatra romanească	Transylvania	1949-1950
Echinox	Transylvania	1973-1974, 1986
Adevărul de Cluj	Transylvania	2003, 2012
Făclia de Cluj	Transylvania	2004, 2008
Bihoreanul	Transylvania	2015

According to Paula Diaconescu [1974], the evolution of the Romanian press, on which we can add a series of significant Romanian contributions [Coman, 1997; 1999; Petcu, 2000; Drăgan, 1996; Szabo, 1999; Ioncioaia, 2000] practically synchronises with the evolution of the modern literary Romanian language, which she divides in the following periods:

a) *the paşoptist period*[44] (1830–1860), focused on the pursuit of the unification and modernization of the literary language;

b) *the period following the Union* (1860–1867–1875), characterised by the imposing of the Latinisms and the search for solutions in view of the standardization of the language;

c) *the great classics period* (1875–1900), marked by individual styles of the contemporary writers, producing the consolidation of the literary language.

However, it should be noticed that the attempts to classify the main moments of the

[44] Considered the beginning of our cultural modernity.

evolution of the Romanian press are relative and conventional. Such an attempt is hindered by clear differences in the consideration of the years that mark the borders between different periods [Chendi, 1900; Hodos & Ionescu, 1913; Lupu *et al.*, 1966/1972; Hangiu, 1987; Preda, 1992; Stoichițoiu, 1992; Răduică & Răduică, 1995; Iliescu, 1999; Stoichițoiu-Ichim, 1996].

Regardless of the staging criteria, the year 1829 remains one of reference for the beginning of the Romanian press. It is then when two significant publications appear: *Curierul românesc*, under the management of Ion Heliade Rădulescu (Bucharest) and *Albina românească* (Iași), with Gheorghe Asachi as the main editor. Only few years later, *Gazeta de Transilvania* appears (1838), printed by Gheorghe Barițiu. It is the moment when we can talk about a Romanian press, in the usual sense of the term, which means relatively constant periodicity, present-day information, consistency and, moreover, an audience eager to inform itself. The process of the evolution of the Romanian press is divided by Ilincan [2010] n more periods, confirmed also by the sociological study of Zamfir & Filipescu [2015] , which focuses only the years 1900-2010. As will be seen, this periodization will be identified in this thesis by statistical techniques, with a tolerance of ±5-7 years (see Chapter 7):

(1) 1829-1860 – which we describe as *the transition period* to the rebirth of the Romanian culture;

(2) 1860-1900 – *the economical evolution period*;

(3) 1900-1947 – *the monarchy period, also that of the developing capitalism in Romania*;

(4) 1947-1989 – *the communist period*;

(5) 1990-2010 – *the post-communist period (the second transition).*

The transition period includes *Curierul românesc*, which, together with *Albina românească*, can be considered the veterans of the Romanian press. It was the first newspaper published in the Romanian language, printed in Bucharest under the management of Ion Heliade Rădulescu. It is the first gazette with a constant periodicity, having a long live, with only a short interruption (8/20 Apr. 1829 – 12 Dec. 1859), with a varied content (administrative texts, political and military news, articles and directed notes, containing elementary notions of History, Geography, Business, Economics, Industry, etc.

Albina românească[45] is the first newspaper in the Romanian language in Moldavia, printed in the beginning with Cyrillic characters. It was founded by Gheorghe Asachi on 1829 in Iași, with a bi-monthly periodicity (1 June 1829 - 24 November 1858, with two breaks: 8 January

[45] Răduică, Georgeta, Răduică, Nicolin. *Dicționarul presei românești (1731-1918)*, Ed. Științifică, București, 1995, p. 5. In the period 9 January 1850 – 24 November 1858, *Albina românească* becomes *Gazeta de Moldavia*; Asachi's publication being continued by *Patria*, "political and literary paper" (Iași, 27 November 1858 – 5 October 1859), edited by Constantin Gane.

- 7 December 1833 and 30 June 1835 - 3 January 1837).

Gazeta de Transilvania (12 Mar. 1838-1944 and 1990-26 Jun. 2009) is the first political newspaper in Transylvania, printed in the beginning by George Barițiu in Brașov. It adopted the semi Cyrillic system in 1844. Between 3 Jan. 1849 and 1 Dec. 1849 it appears by the name of *Gazeta transilvană*, then *Gazeta Transilvaniei*. It appeared under the same name over a time, having a fundamental political influence during the 1848 Revolution, when it fought for national unity of the Romanians from both sides of the Carpathians and for abolishing the serfdom in Țara Românească (Wallachia) - and Moldova.

Foaia pentru Minte, Anima și Literatura[46] is a weekly social and literary publication, which appeared on the 2nd of July 1838 in Brașov, at the initiative of George Barițiu, and continued to exist until 1865. It was a literary appendix of *Gazeta de Transilvania* and used to hold in its pages topics related to individuality assertion through language and culture. Initially, it was printed with Cyrillic characters (only in the title pages they used mixt characters), but starting from 1840 the magazine adopted the semi Cyrillic system.

On the 4th of January 1846 the weekly religious, political and literary gazette (*Gazeta Beserecésca, Politica, e literaria*), *Organulu Luminarei*[47] appears in Blaj from the initiative and under the management of Timotei Cipariu, transformed during the 1848 Revolution in *Organul național*. The publication will appear till 29 September 1848, being the first gazette printed entirely in Latin characters.

Telegraful Român is a project of the first Romanian metropolitan of Ardeal, Andrei Șaguna, and appears in Sibiu at 3 January 1853, under the editorial office of Aaron Florian, declaring itself "for the work of God and of people". Although the title of the bi-weekly gazette was in Latin characters, its body was printed in Cyrillic till 1863. The running life and regularity of this Romanian national journal is remarkable. From the year 1867, for a short period, it becomes *Telegrafulu Romanu*.

The period of economical evolution includes also magazines from the previous period, on which other famous magazines are added. *Familia*, which appears in Budapesta at the initiative of Iosif Vulcan to cultivate the Romanian language in Transylvania, becomes the most important Romanian magazine in Hungary from the second half of the 20th century. From 1880, the *Familia* magazine will appear in Oradea, the place where it is still edited and printed.

The magazine *Convorbiri literare* appears twice a month in Iași at 1 March 1867-1886. It

[46] More about the role of *Foaie pentru Minte, Anima și Literatura* publication in Hangiu Ion, *Dicționarul presei literare românești: (1790-2000*), Ed. Institutului Cultural Român Bucureşti, 2004.
[47] Idem.

is considered the press body of Junimists, having Titu Maiorescu as the spiritual leader. Under the editorial office of Iacob Negruzzi, it appears between 1886-1944 in Bucharest. Nowadays the magazine remains as representative in the Romanian literary field as it used to be.

The magazine *Transilvania*[48] is also connected to George Bariţiu, being first published on 1 January 1868 in Braşov. It is a bimonthly magazine founded as a body of the Transylvanian Association for Romanian Literature and Romanian people (ASTRA). Originally called *Transilvani'a. Foi'a Asociatiunei transilvane pentru literatur'a roman'a si cultur'a poporului romanu*. Interrupted during 1879-1880, the magazine will appear again in Sibiu, at 1 January 1881, Bariţiu resigning in 1871 but returning in the position of director in 1888 for another year.

Federaţiunea[49] is a political, literary, commercial and economic journal (*Diurnalu politicu, literariu, comercialu si economic*), founded by Alexandru Romanu in Pest, fluctuating as periodicity in three time periods: four times a week (15 January to 28 December 1868), three times a week (1 January 1869 – 27 Dec. 1872), twice per week (31 December 1872 – 12 Mar. 1876).

Gura Satului[50] appears at the initiative of Joseph Vulcan (which, until 1902, is also editor once a week) in Pest (5 Jan. 1867 to Feb. 1871) and in Gherla (1 Jan. to 22 Mar. 1881), its message being satirical. On a monthly basis, it appears in Arad (7 March 1871 – 30 September 1879; Series II: 12 January 1901 – 25 April 1903). Between 1863 – 1867 it appears with the title *Umoristul* (Series I), in 1870 it has the subtitle *Organ glumeţ, social-politic*, and in 1900 – *Ziar umoristic*.

6.1.2 Language particularities of the Romanian journalistic

In this section we present some of the language deviations during the monitored period, named "features of journalistic language" [Gîfu, 2015c], having as a basis of research a number of text fragments belonging to publications that have appeared in the three monitored Romanian historical regions. We compare our conclusions with other studies that reflect the lexical and ortographic differences in the Romanian language in different periods of time [Şuteu, 1974; Hristea, 1981].

[48] Bucur Ţincu, Transilvania (1868-1900). In: Ovidiu Papadima, *Reviste literare româneşti din ultimele decenii ale secolului al XIX-lea*, Ed. Academiei Republicii Socialiste România, Bucureşti, 1974, p. 24
[49] More in Ion Hangiu, *Dicţionarul presei literare româneşti: (1790-2000*), Ed. Institutului Cultural Român Bucureşti, 2004.
[50] Idem.

We will exemplify the linguistical differences with some fragments of text.

A fragment from *Albina românească* (no. 46, 12 Jun. 1832): *nerămăindu-i alt doru decât a fi, după cuviință prețuit a ei zilos, și îngăduite ale sale neagiunsuri.*

```
nerămăindu-i = nerămânându-i
doru = dor
zilos = zel
neagiunsuri = neajunsuri
```

A fragment from *Organulu Luminarei* (no. 44, 1 Nov. 1847): *Ateismulu afectatu la aristocrati si cei de frunte, la barbatii, carii gubernéza statulu...*

```
Ateismulu = Ateismul
afectatu = afectat
carii = care
gubernéza = guverneaza
statulu = statul
```

A fragment from *Foaia pentru Minte, Anima și Literatura* (no. 41, 13 Oct. 1854): *... averi ale romaniloru care prin subjugarile facute se totu adaugeau, au introdusu luxulu si pomp'a in vieti'a loru familiare.*

```
romaniloru = românilor
totu = tot
adaugeau = adăugau
introdusu = introdus
luxulu = luxul
pomp'a = pompa
in vieti'a = în viața
loru = lor
familiare = familială
```

A fragment from *Convorbiri literare* (no. 1, 1 Mar. 1867): *Începutul este modestu, dar redacţiunea sperează că va pute da în curendu acestei foi proporţiuni ….*

```
începutu = începutul
modestu = modest
redacțiunea = redacția
sperează = speră
pute = putea
curendu = curând
proporțiuni = proporții
```

A fragment from *Federațiunea* (no. 9, 9/31 Dec. 1868): *Apropiandu-se inceputulu anului 1869, rogâmu pre doritorii de a avè acestu diurnal…*

```
Apropiandu-se = Apropiindu-se
Inceputulu = începutul
Rogâmu = rugăm
pre = pe
avè = avea
acestu = acest
diurnal = jurnal
```

A fragment from *Gura Satului* (no. 20, 7/9 Jun. 1869): *Unu jurnalu ungurescu scrie, câ comunitatea bisericésca unita din Lugosiu…*

```
Unu = Un
jurnalu = jurnal
ungurescu = unguresc
câ = că
bisericésca = bisericească
Lugosiu = Lugoj
```

A fragment from *Transilvania* (no. 21, 1 Nov. 1873): *Vocea tempului de astadi striga inainte, semnulu lui e progresulu,…*

```
tempului = timpului
astadi = astazi
```

semnulu = semnul
progresulu = progresul

A fragment from *Familia* (no. 25, 19 Jun. 1877): *Literatur'a in generalu este cunoscinti'a adânca a regulelor diverselor genuri de compositiuni literarie, in prosa séu in versuri.*

Literatur'a = literatura
generalu = general
cunoscinti'a = cunoașterea
regulelor = regulilor
compositiuni = compoziționi
literarie = literare
prosa = proză
séu = sau

A fragment from *Telegraful Roman* (no. 66, 24 Nov. 1888): *In ședința Camereĭ de Sâmbăta trecută, cuconul Iancu Trece-Gârlă, represintantul sătenilor de la Buzeŭ….*

Camereĭ = Camerei
cuconul => courtesy form, almost not used anymore, synonym with (voievod, domnitor, principe) or copil de boier sau de domn → (En. *prince*);
represintantul = reprezentantul
Buzeŭ = Buzău

Out of the fragments exemplified above we can underline the fact that the differences from the present standard are, often, of phonetics, marking the archaic and regional difference (e.g. adaugeau = adăugau, introdusu = introdus, séu = său), reflected in the spelling of that time. An important landmark in the evolution of diacritical signs is Petru Maior's work [1819], *Orthographia romana sive latino-valachica, una cum clavi*, which presents a complete orthographic system, based also on the etymological principle, but moderated with phonetic rules for easier reading. (e.g.: micả = mică, cảntecu = cîntec, *scii* = *ştii*).

Ex. 1. the use of the apostrophe notes phonological realities from the familiar, casual,

popular or regional speaking: vieti'a = viața, Literatur'a = literatura, cunoscinti'a = cunoașterea; prosa = proză.

Ex. 2. morphological differences (regulelor = regulilor), the plural modification: regule => reguli).

Also, final, breathed, u was specific of the old Romanian language. In the present Romanian language we meet two situations which explain the evolution of this final vowel: the unsounded masculine definite article (calu = calul), and in final positions of participle (făcutu = făcut).

It is also worth noticing some disappeared words or words that are about to disappear:

- cuconul[51] => courtesy form, synonym with domnul (voievod, domnitor, principe) or copil de boier sau de domn → (En. *prince*);

benișel => guler → (En. *collar*);

feledreș => catifea → (En. *velvet*);

scuiburile => cuiburile → (En. *nests*);

vivate => ovații → (En. *ovations*);

chizeșluim => garantăm → (En. *guarantee*), etc.

- the editing of some words in the semi-Cyrillic system, known as the transitional one: gubernéza = guvernează → (En. *govern*). Here, the letter b (Cyrillic alphabet) becomes v (Latin alphabet) through transliteration.

By 1840, the press has contributed to fix unique rules for Romanian literary language by promoting in the widest circles of readers the political, administrative, scientific terminology etc., consisting in the adaptation of neologisms and the gradual removing of some phonemes, old grammatical particularities etc. Between 1840 and 1860 we can see the relinquishment of the official tone, increasing the number of elements borrowed from other styles. After the 60s, the journalistic style is more clearly defined based on the raise of literary language norms, a natural consequence if we think at the nineteenth century newspaper editors' prestige (Andrei Șaguna, Gheorghe Barițiu, Timotei Cipariu, Andrei Șaguna, Iosif Vulcan, etc.). Gradually, the regional or popular elements disappear, outlining the process of language and culture unification.

6.2 The stages of Bessarabian written press between 1867-2015

[51] http://dexonline.net/definitie-cucon

We may say that the actual level of understanding the Bessarabian press phenomenon is still subjective. A reason for this is the lack of a thorough research [Kulikovski, 1957] aimed to describe impartially the characteristics of Bessarabian publishing phenomenon after 1812, when Bessarabia was detached from Romania.

Our computational approach opens up ways for a rigorous research of the Bessarabian journalistic phenomenon in contrast with the Romanian phenomenon, in order to distinguish the dyanamics of similarities and differences of the two variaties of language on the stated time period.

6.2.1 A brief history of Bessarabian newspapers. The corpus description

In this section we will briefly describe the important events (*e.g.*, the appearances of some publications) identified on the acquired corpus (Table 6.2), preparing thus the analysis of the written Bessarabian language that will be the subject of the next chapter.

Table 6.2 The collection of Bessarabian newspapers included in our PhD corpus

Publication	Year(s)	Alphabet
Cuvânt Dimitrie Balica	1848	Cyrillic transl. in Latin
Curierul	1876, 1918, 1932, 1990, 1991	Latin
Candela	1890	Latin
Deșteptarea	1896, 1989, 1990, 1991	Latin
Cuvânt moldovenesc	1913	Latin
Glasul Basarabiei	1913, 1914	Latin
Ardealul	1918	Latin
Basarabia	1918, 1919, 1929, 1932, 1932, 1938, 1941	Latin
România nouă	1918	Latin
Sfatul țării	1918, 1920, 1932	Latin
Democratul Basarabiei	1919	Latin
Luminătorul	1919, 1923-1925, 1928-1930, 1933-1936, 1939, 1941	Latin
Universul	1920	Latin
Dreptatea	1922, 1926	Latin
Solidaritatea	1927, 1930, 1932, 1934, 1940	Latin
Basarabia Chișinăului	1931	Latin
Adevărul	1933	Latin
Buletinul Arhiepiscopiei Chișinăului	1942, 1944	Latin
Viata economică din Bălți	1943	Latin
Discurs Onufrie Andrus RSSM	1945	Latin
Ethos	1945	Latin

Sania	1955	Cyrillic transl. in Latin
Tinerimea Moldovei	1962	Cyrillic transl. in Latin
Moldova socialistă	1967	Cyrillic transl. in Latin
Rev. Univ. Chișinău	1968	Cyrillic transl. in Latin
Cuvînt Ion Druță	1970	Cyrillic transl. in Latin
Povara bunătății noastre	1970	Cyrillic transl. in Latin
Viața satului	1975	Cyrillic transl. in Latin
Pedagogul	1980	Cyrillic transl. in Latin
Femeia Moldovei	1986	Cyrillic transl. in Latin
Literatură și artă	1989, 1990, 1991, 1998, 2015	Latin
Jurnal	2009-2010	Latin
Contrafort	2013-2014	Latin
Jurnal de Chișinău	2013	Latin
Moldova suverană	2014	Latin
Ziarul de gardă	2014	Latin

1. The tsarist occupation phase (1812-1917). Bessarabia was named "the tsarist Bessarabia" after its annexation to the Russian Empire in 1812, following the Bucharest Peace in 1812. The literature records reference works about the 19th century Bessarabia [Mihail & Mihail, 1993; Bobeică, 1993; Ciobanu, 1992a; Avramescu & Avramescu, 1996/2001 ș.a.m.d]. Perhaps the most representative contribution remains Cobanu's historiographical synthesis, which offers an overview about the Romanian periodicals existing in Bessarabia between 1867 and 1917. There are appreciations on the Marian's study, even if it is incomplete as value [Marian, 1932]. Furthermore, in [Ciobanu, 1992b] Ciobanu manages to draw some notes about some publications written in Russian that captures the nationalist spirit of Romanian Bassarabians (Alexandru Sturdza, George Gore's published work, Alexis Nacco, John Halippa, Alexei Mateevici, Leon Casso).

Remarkable too is Boldur's work [1940:23], recording the fact that during 1812-1880, 116 Romanian prints appeared in Bessarabia, and 114 between 1881-1918, with themes focused mainly on folklore and religion.

Moreover, according to the priest Paul Mihailovici, who reviewed the press of the Bessarabian occupied territory, 480 cataloged editions appeared there, implying that the total Romanian prints in Bassarabia was close to a thousand [Boldur, 1943:72]. Interested about the Bassarabian media in the period preceeding the 1918 great Union was also Onisifor Ghibu [1926], one of the promoters of Bassarabians revival of national consciousness. His conclusion is that the success of the great union of 1918 is in great part due to the Romanian publications in this area.

According to Boldur, this period can be divided as such: *before 1870*, when Romanian

publications appeared without too many obstacles (e.g., the clerical magazine *Bessarabskia Eparhialnâia Vedomosti* (*Buletinul eparhiei basarabene*), both in Russian and in Romanian, between 1 July 1867-1904; „filișoarele" Frăţiei *Naşterea lui Christos* from Chișinău, organization established in 1899 close to the Nativity Cathedral, and the Romanian newspaper *Mesagerul Basarabiei* in Chișinău, published in 1884); *the oppression era* (1870-1905), of almost absolute Romanian language silence, since only in 1903 appears *Bessarabskaia Jizni*, the first democratic newspaper, that resisted until 1918; and *the revival era from 1905*, when Romanian publications were allowed again (e.g., *Junimea Literară* in Cernăuţi, printed between 1 January 1904 till 1914 and *Buletinul eparhiei basarabene* that appeared between 1905-1917). Incidentally, the first Russian revolution causes also the first reactions in Bessarabia, in 1905 appearing in Switzerland the first bassarabian newspaper *Bassarabia*, founded by Zamfir Arbore. One year later, on 24 May 1906, Emmanuel Gavriliţă's *Basarabia* is printed in Chișinău, with a revolutionary and nationalist blast. Two years later, the Church magazine appears, then *Luminătorul* (1 January 1908-1944) – the only Moldavian periodical publication that animated the Romanian language writers at the time.

2. 1918-1940 – the testing phase

Bessarabia's interwar media history segment is recorded in the works of Șpac [2002; 2006], who gives significant information selected from the collection of *Viaţa Basarabiei* (1932-1944) and *Arhivele Basarabiei* (1929-1938). In addition, collective works focused on valuing of patrimony collections are reminded in [Danilov, 2005].

An approach of the Bessarabian reality paradoxes we find in [Mâțu, 1930]. The situation of the Romanian language press was more problematic than in the previous period. Basically, until 1940, in Chișinău are edited only very few periodicals: *Viața Basarabiei* – the most important interwar publication of Bessarabia; *Gazeta Basarabiei*; *Furnica*, intended for ploughmen; *Basarabia Economică* – a magazine displaying socio-economic studies over Bessarabia; *Itinerarul*; *Ghiocelul* etc.

The Romanian written press is also present in Orhei, where *Speranța* appears (from 1935), Bălți – *Buletinul Asociației Pedagogice din județul Bălți*, Soroca – *Solidaritatea*, a newsletter of the Association of Teachers, appeared in 1923, and in Southern Bessarabia, where *Bugeacul*, a promoter of cultural regionalism, was edited.

3. 1940-1945 – the Bessarabian's press during the Second World War

After the occupation of Bessarabia by the Red Army, in June 1940, and its transformation into a Soviet Republic, several institutions were transferred out of Chișinău (e.g. the State Publishing House in Tiraspol). The media remained only with an ideological role addressing the masses. "In that period, the linguistic argument of Serghievski, who claimed that there appeared a new Eastern Roman language, Moldavian, independent from the Romanian language, prevailed." [Constantin, 1994:63]. It is a period of suffocation for the Romanian cultural life, detailed in [Kulicovski, 2002]. The objectives of desnationalization where prioritized, with a huge widespread of Russian newspapers for propaganda purposes. In 1940, Chișinău become an administrative center of the region Nistru with only 4 districts: Orhei, Lăpușna, Tighina și Cetatea Albă [Țurcanu, 2004].

The short period of Bessarabia's and North Bucovina's release from occupation (sept. 1941 – mar. 1944) and their affiliation to Romania coincides with a regain of press refreshness. *Luminătorul*, *Cuvînt Moldovenesc*, the weekly *Raza* etc., appear again, propagating messages of Romanian national values. In the next chapter we will highlight the main topics of the press, revealed by statistical means. [Gîfu, 2016a, d].

Luminătorul and *Misionarul* were publication of a clerical profile, appearing monthly in the period 1941-1944, the last one as an Organ of the Archdiocese of Chișinău [Grossu, 2003]. Other religious publications of the time were: *Biserica basarabeană*, *Buletinul Arhiepiscopiei Chişinăului*, *Buletinul oficial al provinciei Basarabia*.

They were also clerical publications in exile (*e.g. Raza Basarabiei*, between 1940-1944, edited in Bucharest) of a cultural nature and national propaganda use. This should be considered to follow *Bessarabia* (1931-1940), an independent political newspaper with a similar orientation. A weekly publication that helped to maintain the national consciousness was *Basarabia Literară* (April 1942 – February 1944), a supplement of *Basarabia*.

There is worth mentioning also other publications, such as *Viata Basarabiei*, a magazine published in Bucharest, supressed in Chișinău in 1940, *Cuget moldovenesc*, in Iași, *Cuvînt Moldovenesc*, appearing weekly in Chișinău in the period 1941-1944, *Solidaritatea*, the only pedagogical magazine, which reappears between 1942-1944. Also, starting in 1942, the "M. Kogălniceanu" Cultural Society publishes the magazine *Luminile tineretului*, which helped young literary talents.

4. Between science and propaganda (1944-1989)

The period 1944-1989 was often described as the "Developped Socialism", when in Bessarabia nothing was published in the Latin script (except for the war years). One could discern two research directions regarding the historiography of the Bessarabian press [Danilov, 2012]:

1. The 1960 – 1970 period, when all publications made up the so called "Party" press, actually a Soviet press, because was written in Russian;

2. The 1970 – 1980 period, when there is a re-focuss towards the Moldavian – Russian – Ukrainian relationships, as well as on the role of the Russian literature and press in the consolidation of the Bessarabian journalisms.

There are a few notable researchers who tackled with the local Bessarabian press during the above mentioned period. Among these, Dumitru Coval [1991; 1992], for his reference monographic studies and Efim Levit [1979], etc.

5. The valorization of sources after 1989

The Bessarabian literature of after 1989 was recorded in a number of studies and articles that appeared in specialised magazines. In 1991, the Cultural Society *Mitropolitul Varlaam* was founded. One year later, within its frames, the *Luminătorul* magazine reappeared, holding its reputation as the longest magazine ever published. One of the most important documents regarding the Bessarabian Press is a dictionary called *Catalogul Presei Basarabene*, which contains newspapers and magazines on the one hand, and newsletters, yearbooks, calendars and loose sheets on the other.

In addition to that, in the recent historiographical scenery of the Bessarabian press there is yet another essential appearance: the – Thematic Illustrative Almanac, since 2003, consisting of a compilation of historically oriented subjects and "investigational" articles, published for the last five years in the *Novaja Vremea* magazine (Новоя Время) (more than 200 articles "of investigation" with historical topic).

6.2.2 The Bessarabian journalistic language particularities

Here it is a fragment from *Cuvântul introductiv,* signed by the traveller Dimitrie Balica, 1848: *De cătrŭ al mieu rău noroc aruncat, găsăndu-sŭ intre-al Besarabiei pămiant trecut acum 12 ani.*

cătrŭ = către

mieu = meu

găsăndu-sŭ = găsindu-se

intre-al = într-al

Besarabiei = Basarabiei

pămiant = pământ

Here it is a fragment from *Candela,* 1 December 1890: *Direcţiunea moşiilor se aplecară şi la amploĭaţiĭ dela aşeḑămintele montane...*

Decemvrie = decembrie

Direcţiunea = direcţia

amploĭaţiĭ = angajaţii

dela = de la

aşeḑămintele = aşezămintele

Here it is a fragment from *Plugarul Roşu* (Плугарул Рош, transliterated Plugarul Roș), 1920: *De-amu v-o două luni di zăli, dicând Plugarul Roş îşi lunjeşte discusîia dispri orfografia moldovenească, mai întîi trebui di spus cî sfada merji nu dispari limba moldovenească, adicî dispri sămnuirea suniţilor cari sînt în limba jii moldovenească.*

di = de

zăli = zile

dicând = de când

discusîia = discuţia

dispri = despre

moldovenească = moldovenească

cî = că

merji = merge

dispari = dispare

adicî = adică

sămnuirea = semnarea

cari = care

limba jii = limbaj

From the fragments exemplified above we can underline the fact that the differences from

the present standard language are, often, of a phonological origin, which indicates the archaic variants of writing (e.g. `limba jii = limbaj`), specifically in the Moldavian idiom (*e.g.* `di = de, dispri = despre, cî = că)`, reflected in the spelling of that time, orthographical differences, etc. Also, the palatization phenomenon, occuring in Moldavian words ending in `i/î` is specific to this idiom.

The re-introduction of the Latin alphabet for the Romanian writing, officially adopted in 1881, marks the adaptation of the Romanian sounds. The distinctiveness of the Latin vocalism is reflected either through the preservation of the different short vowels from the initial syllable in opposition to the others (e.g. *genus*[52]: *geneses* > *generis* → (Ro. *gen*) → (En. *gender*) *or optumus*: *optimus* → (Ro. optim*)* → (En. *optimal*)), or through the tendency of reduction of the articulation (e.g. *treyes*: *trēs* → (Ro. *trei*) → (En. *three*) *or koopia*: *cōpia* → (Ro. *copia*) → (En. *copy*)).

Furthermore, during 1945-1989, the so-called "developed socialism" period the Latin spelling was banned. Thus, to make use of documents of that period, our pre-processing machinery had to include one more module, the transliteration, which triggered a certain loss of constancy with regard to the letter of the initial text, an issue signalled also in [Munteanu, 2008]. Along the period 1945-1989 nothing was published in Bessarabia with a Latin spelling. The press was written in the Russian alphabet, out of which three letters were excluded. Later on, a specific letter was added to better reflect the Romanian spelling (called "`j`" with diacritics), which rendered the sounds of "`ge`" or "`gi`". The vocabulary and the ortography were very different till 1951 and after. [Boian *et al.*, 2014; Petic & Gîfu, 2014].

6.3 Conclusions

This chapter summarises the different phases in the evolution of the Romanian language as it was manifested in the written press over the Romanian and Bessarabian territories. The discussed periods cover the interval 1829-2015 for Romania and 1867-2015 for Bessarabia. Very small extracts also display samples of language in view of preparing the analysis and the discussion following it of the next chapter. As will be seen, the monitored and representative publications will help to build a lexicon of words and phrases that will be correlated with the existing Romanian language resources: DEX-online, Ro-WordNet, and e-DTLR.

In the intervals when the press was not politically constrained, the gazettes often

[52] http://latindictionary.wikidot.com/noun:genus (The Latin-English Dictionary).

functioned as spear heads, reflecting the cultural values and promoting the advanced linguistic tendencies of the epochs. On the contrary, in the totalitarianism intervals, the press being commanded by the state, it is to believe that the language they cultivated was to a certain extend artificial. We believe thus that the historical conditions in the two territories under analysis, at different moments, favored either the closeness of the two varieties of language or, on the contrary, their diversity. These natural oscillations presumably make of our corpus a good experimental field for a computer modelling, provided enough data are available. As will be commented in the next chapter, the extension of the corpus remains a priority because the accuracy of the data obtained through machine learning depends on the dimension of the analyzed data base.

7. AN EVALUATION OF NEWSPAPERS LANGUAGE IN DIACHRONY

In written form (e.g. newspapers), language serves as an eternal witness of knowledge from one generation to the next, especially when the separation is influenced by territorial borders. This chapter describes a methodology of diachronic investigation on news corpora, which determines the degree of similarity between cognate languages. We talk about a large corpus (over 3 million lexical tokens), chronologically ordered since the second decade of the 19th century, developed and structured in four independent collections of newspapers corresponding to Moldavia, Wallachia, Transylvania and Bessarabia. Nowadays the first three regions form Romania. Bessarabia was a part of Romania until 1812 and between 1918 and 1940.

We think this model of language could be useful both for scientific objectives (for exploring the nature of cognate languages) and for applicative objectives (for enabling effective language similarity analysis using statistical methods).

7.1 Architecture of the diachronic study

In this section, we describe the architecture on which we have conducted our diachronic study (for short, diachronic architecture, see Figure 7.1), structured in 4 modules:

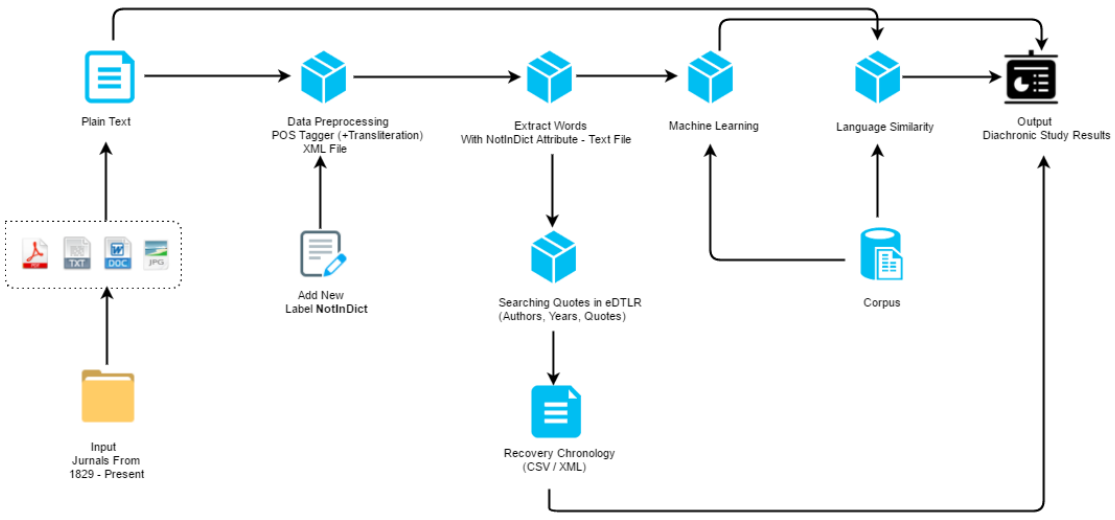


Figure 7.1 Diachronic Architecture

7.1.1 Processing the corpus

This phase (Figure 7.2) consists in cleaning the text received as input. We must take into consideration two types of inputs: plain text and PDF, out of which raw text is beeing extracted, by using some tools (e.g. a Java PDF Library - Apache PDFBox[53]).

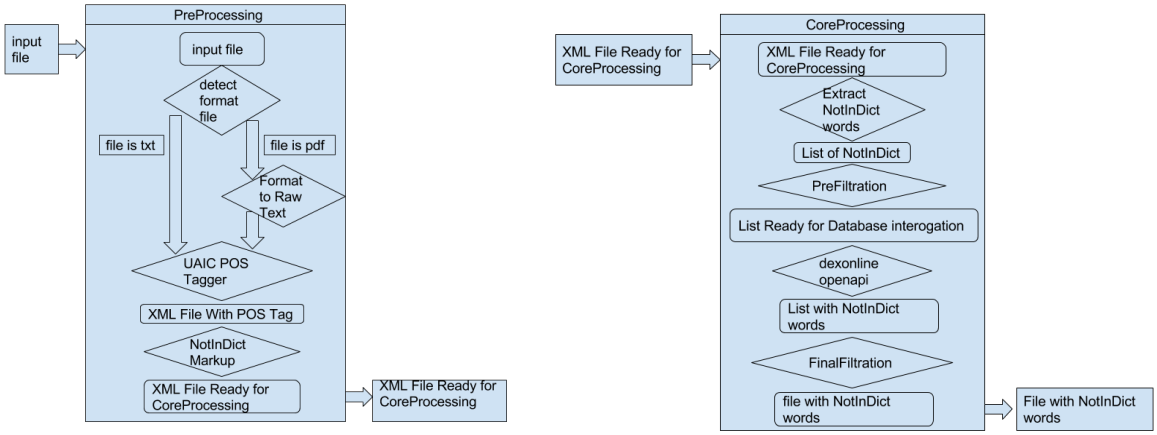


Figure 7.2 Processing the corpus

At this level, a sequence of steps follows:

1. *Extracting the content of input*.
   Our system compatible files are TXT and PDF.
2. *Improving the format file* using UTF-8 encoding for Romanian diacritics recognition.
3. *Cyrillic-Latin transliteration*

Cyrillic-Latin transliteration was a necessary step in preparing our corpus[54], given that, between 1944-1989, in Bessarabia, nothing was published in the Latin script (except for the war years). For the Bessarabian corpus representativeness, we considered the publications of the mentioned period, with the support of the Academy of Sciences of Chișinău. They were scanned, OCR-ized [see Moruz *et al.*, 2012] and transliterated[55] using a number of transliteration

[53] https://pdfbox.apache.org/
[54] In contrast to Bessarabia, in Romania, the publications were in both Latin and Cyrillic script, and in the early 1860s, the Latin alphabet was formally adopted. In [Haja et al., 2005] we find that the first work in which the multiple etymological principle was applied is *Dicționarul limbii române* (the Romanian Language Dictionary), written between 1871-1876 by A. T. Laurian & I. C. Massim. It was a criticized work in [Densușianu, 1901, 1938].
[55] We mention the existence of the SR ISO 9:1997 standard, which establishes a system of transliteration from Cyrillic to Latin characters.

heuristics, already tested on the existent Cyrillic-Latin lexicon[56] [Boian *et al.*, 2013].

In Figure 7.3 (a), (b) a text in Cyrillic and in Latin after transliteration can be visualized. According to [Boian *et al.*, 2013/2014; Petic & Gîfu, 2014], the vocabulary and the spelling were very different until 1951 and even after[57].

Меандреле индиференцей

Пункте де ведере

Савантул франчез Б. Паскал а констатат ын «Кужетэриле» сале, кэ презентул ну есте нич одатэ скопул ностру. Трекутул ши презентул не сынт мижлоаче; виторул есте цинта ноастрэ. Ачесте кувинте ар требуи сэ фие о кэлэузэ пентру фиекаре тынэр. Уний, ынсэ, трэеск доар ку трекутул ши ку презентул, неавынд ун идеал конкрет пентру виитор. Се ынтымплэ аша, де челе май мулте орь, дин кауза инфлюенцей негативе а медиулуй. Ын привинца ачаста ам авут о дискуцие ку Михаил Хангану, студент ын анул III, филоложие.

— Ам венит ла университате ку ун идеал маре — сэ девин педагог бун. Кредям кэ студенций сынт чей май весель, чей май романтичь. Ши че кредець? Дин анул ынтый м'ам конвинс кэ ын журул меу сынт фоарте мулць че трэеск доар ку о луме а лор, ынгустэ, фэрэ сэ се интересезе де марile проблеме але вецій.

— Врей сэ зичь, Михаиле, кэ медиул ачеста а авут о инфлуенцэ асупра идеалулуй тэу?

— Десигур. Да маре импортанцэ аре ши фиря студентулуй. Дакэ ел ынтр'адевэр шь-а пус скопул сэ девинэ педагог де трябэ, инфлуенца есте май микэ. Дин ачештя фак парте пуцинь тотуш: И. Телешку ши Е. Мунческу анул III, Л. Белый дин анул IV, Л Притула дин анул V ш. а.

Дар ну тоць сынт енержичь ши деачея мажоритатя студенцилор се молипсеск де индиференца ачаста. Авем студенць кодашь «ыннэскуць», пе каре де ла бун ынчепут ну-й атражэ ынвэцэтура. Дау де доуэ-трей орь ун екзамен ку гындул кэ лекторул ле ва пуне ын челе дин урмэ ун треишор. Аша ажунг пынэ ын анул патру, ши токмай атунч, ла практика педагожикэ, се трезеск... Дар е тырзиор!

— Ай дрептате, Михаил. Зилеле трекуте м'ам ынтылнит ку Константин Мунтяну, ексклус анул трекут дин университате. Ел мь-а дестэйнуит кэ дакэ ар фи фост ексматрикулат ын анул I, поате кэ шь-ар фи гэсит пынэ акума ун рост ын вяцэ, дар аша...

— Ынсэ унеорь се ынтымплэ кэ ун тынэр, кэруя ну и с'а спус ла тимп кэ требуе сэ се апуче сериос де карте, терминэ университатя. Луй и се ынмынязэ диплома де ынвэцэтор, и се ынкрединцязэ копиий пе каре требуе сэ-й едуче, сэ ле формезе густул естетик. Дар че-о сэ-й ынвеце ачеста, дакэ ел ынсушь н'а манифестат интерес фацэ де челе май елементаре проблеме але артей? Пробабил, о сэ ле спунэ елевилор нумай чея че есте ын мануал, кум с'а ынтымплат ку Серафим Козма.

— Васэзикэ, студенций се ымпарт ын доуэ категорий: медиокри ши бунь.

— Ну. Май сынт о парте, каре ла прима ведере пар а фи дестул де бунь: дау екзаменеле пе «4» ши «5», рэспунд систематик ла семинаре, ынсэ аич се искэ алтэ проблемэ. Ей ынвацэ нумай чея че ле рекомандэ професорул, ярмай департе... трэеск ку букия конспектулуй. Ачештя, кред еу, ну сынт ку мулт май бунь декыт ачей медиокри, деоарече ей ау сэ термине студиile ку ачелаш оризонт де куноштинце, дестул де ынгуст. Ам сэ адук ун екземплу фоарте грэитор ын ачастэ привинцэ. Ын фиекаре зи вэд ын сала де лекторэ доуэ студенте (К. К. ши Е. М. дин анул III филоложие). Еле ынвацэ мулт, дар одатэ, кынд ле-ам ынтребат, принтре алтеле, каре е ултима кулежере де версурь а луй Г. Виеру, мь-ау спус ку ун фел де регрет:

— Ну не пря интересэм де чея че не ынтребь. Ну не ажунже тимп сэ ынвацэм мэкар чея че ни се дэ пентру лекций.

— Да, Михаил, паре парадоксал. Динтр'о конворбире ку М. К., студент ын анул II, бинишор вэзут, ам ынцелес, кэ н'а фост нич одатэ ла театру (се стэруе сэ бируе франчеза, дисчиплина ла каре се симте май слаб), немайворбинд сэ фие аскултат кындва ун кончерт.

— Дар везь ту вре-о кале де а-й салва пе студенций де ачастэ натурэ, де а-й фаче сэ тиндэ спре ун идеал?

— Пентру а рэспунде ла ачастэ ынтребаре, ам сэ нумеск ун фапт, каре м'а импресионат мулт. Ын примул ан де студий, кынд ам плекат ла лукру ын колхоз, ынтр'о сарэ мь-а атрас атенция о групэ де студенць, каре мержэау спре пэдуре. Ынтрынд ын ворбэ ку ей, ам афлат кэ секретарул комсомолист ал групей I А. Скутару а хотэрыт сэ организезе о дискуцие ку тема «Че е фрумосул?» Ын журул унуй руг, с'а дискутат пе тема ачаста пынэ пе ла мезул нопций. Тоць ерау синчерь, ышь дэдяу пэреriле, ышь комбэтяу колеяий, кынд ну ерау де акорд ку ей. А фост о сарэ кынд франкеца й-а унит ши май стрынс пе ачешть студенць, кынд а фост пус ынчепутул унеи серий де дискуций интересанте.

Деч, ун рол импортант ыл аре иницiатива секретарулуй комсомолист. Фииндкэ ел ышь куноаште колеяий, поате гэси чева каре сэ-й атрагэ. Секретарул комсомолист, шефул групей, ку конкурсул лекторилор ар требуи сэ организезе дискуций деспре кэрць, деспре филме, деспре диферите проблеме але вецій студенцешть, ын партикулар деспре студенций дин групэ. Ну е казул сэ аштептэм сэ винэ чинева ын групэ ши сэ не факэ «активитате». Ну путем ной, ынвэцэторий де мыне, сэ луэм инициатива..?

— Кред, Михаил, ар фи бине ка бироул комсомолист ал факултэций сэ организезе дискуций де ачест фел пентру тоць студенций де ла факултате. Е де дорит сэ се организезе май дес ынтылнирь ку оамений де артэ, каре сэ не ынтродукэ ын тайнеле процесулуй де креацие, пе каре — сэ фим синчерь! — мулць ну ле ынцележем. Ла факултате астфел де ынтылнирь ау лок, дар фоарте рар ши деокамдатэ нумай ку уний поець тинерь де ла каре ну пря авем че ынвэца. Дакэ де ла ынтылниря ку Г. Водэ, де екземплу, м'ам алес ку мулте гындурь, апой де ла алтеле де ла сфыршитул анулуй де студий трекут...

— Унеорь о инициативэ фрумоасэ а студенцилор есте «ынзбушитэ» ануме де бироул комсомолист. Аша с'а ынтымплат ку конкурсул пентру чя май реушитэ инсигнэ а факултэций. Ау фост институите премиile, ау апэрут унеле десене дестул де реушите ши... ку ачаста с'а терминат.

— Кам трист стау лукруриле... Ши тотуш, крезь кэ тынэрул че терминэ университатя е пасив дин ачесте каузе?

— Ын ынтрегул диалог ам врут сэ сублиниез ачест гынд. Ын орьче каз, май мултэ инициативэ ну не-ар стрика де лок. Че зичь?

— Че зик еу? Сау че зик чититорий? Кред кэ путем дискута: конкрет, ку май мулте сужестий.

Андрей ЦУРКАНУ.

Figure 7.3 (a) Print screen - *Revista Universităţii din Chişinău*, 17 September 1968, in Cyrillic

[56] http://www.math.md/elrr/ - Institute of Mathematics and Computer Science, Moldova, Reusable Resources for Romanian Language Technology, 2003.

[57] Initially, three letters of the Russian alphabet were excluded, then one, specific to the Romanian language, was added (noted "j" with hat), which plays back the sounds “ge” or “gi”.

Meandrele indiferenței «Chișinevschii universitet» 17 septembrie 1968

Savantul francez B. Pascal a constatat în «Cugetările» sale, că prezentul nu este nici odată scopul nostru. Trecutul și prezentul ne sînt mijloace; viitorul este ținta noastră. Aceste cuvinte ar trebui să fie o călăuză pentru fiecare tînăr. Unii, însă, trăesc doar cu trecutul și cu prezentul, neavînd un ideal concret pentru viitor. Se întîmplă așa, de cele mai multe ori, din cauza influenței negative a mediului. În privința aceasta am avut o discuție cu Mihail Hanganu, student în anul III, filologie.

- Am venit la universitate cu un ideal mare — să devin pedagog bun. Credeam că studenții sînt cei mai veseli, cei mai romantici. Și ce credeți? Din anul întîi m-am convins că în jurul meu sînt foarte mulți ce trăesc doar cu o lume a lor, îngustă, fără să se intereseze de marile probleme ale vieții.

- Vrei să zici, Mihaile, că mediul acesta a avut o influență asupra idealului tău?

- Desigur. Da mare importanță are și firea studentului. Dacă el într-adevăr și-a pus scopul să devină pedagog de treabă, influența este mai mică. Din aceștea fac parte puțini totuși: I.Teleșcu și E. Muncescu anul III, P. Belăi din anul IV, L Pritula din anul V ș. a. Dar nu toți sînt energici și de aceea majoritatea studenților se molipsesc de indiferența aceasta. Avem studenți codași «înnăscuți», pe care de la bun început nu-i atrage învățătura. Dau de două-trei ori un examen cu gîndul că lectorul le va pune în cele din urmă un treișor. Așa ajung pînă în anul patru, și tocmai atunci, la practica pedagogică, se trezesc... Dar e tîrzior!

- Ai dreptate, Mihail. Zilele trecute m-am întîlnit cu Constantin Munteanu, exclus anul trecut din universitate. El mi-a destăinuit că dacă ar fi fost exmatriculat în anul I, poate că și-ar fi găsit pînă acuma un rost în viață, dar așa...

- Însă uneori se întîmplă că un tînăr, căruia nu i s-a spus la timp că trebuie să se apuce serios de carte, termină universitatea. Lui i se înmînează diploma de învățător, i se încredințează copiii pe care trebuie să-i educe, să le formeze gustul estetic. Dar ce-o să-i învețe acesta, dacă el însuși n-a manifestat interes față de cele mai elementare probleme ale artei? Probabil, o să le spună elevilor numai ceea ce este în manual, cum s-a întîmplat cu Serafim Cozma.

- Vasăzică, studenții se împart în două categorii: mediocri și buni.

- Nu. Mai sînt o parte, care la prima vedere nap a fi destul de buni: dau examenele pe «4» și «5», răspund sistematic la seminare, însă aici se iscă altă problemă. Ei învață numai ceea ce le recomandă profesorul, iar mai departe... trăesc cu buchia conspectului. Aceștea, cred eu, nu sînt cu mult mai buni decît acei mediocri, deoarece ei au să termine studiile cu același orizont de cunoștințe, destul de îngust. Am să aduc un exemplu foarte grăitor în această privință. În fiecare zi văd în sala de lectură două studente (C. C. și E. M. Din anul III filologie). Ele învață mult, dar odată, cînd le-am întrebat, printre altele, care e ultima culegere de versuri a lui G. Vieru, mi-au spus cu un fel de regret:

- Nu ne prea interesăm de ceea ce ne întrebi. Nu ne ajunge timp să învățăm măcar ceea ce ni se dă pentru lecții.

- Da, Mihail, nape paradoxal. Dintr-o convorbire cu M. C., student în anul II, binișor văzut, am înțeles, că n-a fost nici odată la teatru (se străue să birue franceza, disciplina la care se simte mai slab), nemaivorbind să fie ascultat cîndva un concert.

- Dar vezi tu vre-o cale de a-i salva pe studenții de această natură, de a-i face să tindă spre un ideal?

- Pentru a răspunde la această întrebare, am să numesc un fapt, care m-a impresionat mult. În primul an de studii, cînd am plecat la lucru în colhoz, într-o sară mi-a atras atenția o grupă de studenți, care mergeau spre pădure. Intrînd în vorbă cu ei, am aflat că secretarul comsomolist al grupei A. Scutaru a hotărît să organizeze o discuție cu tema «Ce e frumosul?» În jurul unui rug, s-a discutat pe tema aceasta pînă pe la mezul nopții. Toți erau sinceri, își dădeau părerile, își combăteau colegii, cînd nu erau de acord cu ei. A fost o sară cînd francheța i-a unit și mai strîns pe acești studenți, cînd a fost pus începutul unei serii de discuții interesante.

Deci, un rol important îl are inițiativa secretarului comsomolist. Fiindcă el își cunoaște colegii, poate găsi ceva care să-i atragă. Secretarul comsomolist, șeful grupei, cu concursul lectorilor ar trebui să organizeze discuții despre cărți, despre filme, despre diferite probleme ale vieții studențești, în particular despre studenții din grupă. Nu e cazul să așteptăm să vină cineva în grupă și să ne facă «activitate». Nu putem noi, învățătorii de mâine, să luăm inițiativa..?

- Cred, Mihail, ar fi bine ca biroul comsomolist al facultății să organizeze discuții de acest fel pentru toți studenții de la facultate. E de dorit să se organizeze mai des întâlniri cu oamenii de artă, care să ne introducă în tainele procesului de creație, pe care — să fim sin-ceri! — Mulți nu le înțelegem. La facultate astfel de întâlniri au loc, dar foarte rar și deocamdată numai cu unii poeți tineri de la care nu prea avem ce învăța. Dacă de la întâlnirea cu G. Vodă, de exemplu, m-am ales cu multe gânduri, apoi de la altele de la sfârșitul anului de studii trecut...

- Uneori o inițiativă frumoasă a studenților este «înăbușită» anume de biroul comsomolist. Așa s-a întâmplat cu concursul pentru cea mai reușită insignă a facultății. Au fost instituite premiile, au apărut unele desene destul de reușite și... cu aceasta s-a terminat.

- Cam trist stau lucrurile... Și totuși, crezi că tânărul ce termină universitatea e pasiv din aceste cauze?

- În întregul dialog am vrut să subliniez acest gând. În orice caz, mai multă inițiativă nu ne-ar strica de loc. Ce zici?

- Ce zic eu? Sau ce zic cititorii? Cred că putem discuta: concret, cu mai multe sugestii.

Andrei ȚURCANU

Figure 7.3 (b) Print screen - *Revista Universității din Chișinău*, 17 September 1968, in Latin

In the transliteration process, we have kept the original spelling of words:

Ex: `birue` instead of `biruie` → (EN) `win`
`trăesc` instead of `trăiesc` → (EN) `live`

Moreover, the writing in the Cyrillic script, as is usual in Russian, the phonetic principle is respected.

Ex: `Șexpir → Shakespeare` (in the Russian script: `Шекспир`).
`Ian Sebastian Bah → Jan Sebastian Bach`
`Persel → Purcell`

```
Tochio → Tokyo
```

Here – an exemple in XML: *când într-o bună vreme...* → (EN) *when in a good period of time...*

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?><POS_Output>
<S id="2" offset="8">
    <W LEMMA="când" MSD="Rg" POS="ADVERB" cyr="Кынд" id="2.1"
offset="0">Când</W>
    <W LEMMA="între" MSD="Sp" POS="ADPOSITION" cyr="ынтр'" id="2.2"
offset="5">într-</W>
    <W Case="direct" Gender="feminine" LEMMA="un" MSD="Tifsr"
Number="singular" POS="ARTICLE" Type="indefinite" cyr="о" id="2.3"
offset="10">o</W>
    <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="bun"
MSD="Afpfsrn" Number="singular" POS="ADJECTIVE" cyr="бунэ" id="2.4"
offset="12">bună</W>
    <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="vreme"
MSD="Ncfsrn" Number="singular" POS="NOUN" Type="common" cyr="време"
id="2.5" offset="17">vreme</W>
...
  </S>
```

4. Then, the process continues with: *segmentation*, *tokenization*, *lemmatization*[58], and *part-of-speech (POS) tagging*, using the UAIC POS-Tagger [Simionescu, 2011].
5. *NotInDict Markup* (default by the POS-Tagger).

For instance: ... *mai iantăi acistu al mieu prostatnic poem cătrŭ domnievoastrŭ*. → (approx. EN) ... *firstly this prostatnic poem of mine to you*.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<POS_Output>
<S>
    <W LEMMA="mai" MSD="Rg" POS="ADVERB" id="null.242"
offset="1323">mai</W>
```

[58]Note that from those features is the lemma of each word which will help us to find words in dictionaries.

```
    <W Case="direct" Definiteness="no" EXTRA="NotInDict"
Gender="masculine" LEMMA="iantăi" MSD="Afpmprn" Number="plural"
POS="ADJECTIVE" id="null.243" offset="1327">iantăi</W>
    <W Case="direct" Definiteness="no" EXTRA="NotInDict"
Gender="masculine" LEMMA="acistu" MSD="Afpmsrn"
Number="singular" POS="ADJECTIVE" id="null.244"
offset="1334">acistu</W>
    <W Case="direct" Gender="masculine" LEMMA="al" MSD="Tsmsr"
Number="singular" POS="ARTICLE" Type="possessive" id="null.245"
offset="1341">al</W>
    <W EXTRA="NotInDict" LEMMA="mieu" MSD="Y" POS="ABBREVIATION"
id="null.246" offset="1344">mieu</W>
    <W Case="direct" Definiteness="no" EXTRA="NotInDict"
Gender="masculine" LEMMA="prostatnic" MSD="Afpmsrn"
Number="singular" POS="ADJECTIVE" id="null.247"
offset="1349">prostatnic</W>
    <W Case="direct" Definiteness="no" Gender="masculine"
LEMMA="poem" MSD="Ncmsrn" Number="singular" POS="NOUN"
Type="common" id="null.248" offset="1360">poem</W>
    <W Case="direct" Definiteness="no" EXTRA="NotInDict"
Gender="masculine" LEMMA="către" MSD="Afpmsrn" Number="singular"
POS="ADJECTIVE" id="null.249" offset="1365">către</W>
    <W Case="direct" Definiteness="no" EXTRA="NotInDict"
Gender="masculine" LEMMA="domnievoastrŭ" MSD="Afpmsrn"
Number="singular" POS="ADJECTIVE" id="null.250"
offset="1370">domnievoastrŭ</W>
</S>
</POS_Output>
```

The result of the processing stage is an XML file. Moreover, we apply GGS[59] grammar rules over the previous file. The GGS rules practically help to disambiguate the hyphen, in order to distinguish a hyphenation at the end of a row from a compound word.

[59] Graphical Grammar Studio (GGS) is a tool used for developing grammar rules, created by R. Simionescu [2016] at UAIC-FII. As input, it supports XML files that include markings at the sentence and word levels (typically <S> and <W>).

For instance:

- inter-word hyphen: `să-n`
- end of a row hyphen: `gus-toasŭ`

As a result of this process, an XML file is obtained. The next steps continue the process of identification and classification, by using some external linguistic resources (e.g. DEX-online).

6. *Extracting words with the extra markup NotInDict* using JDOM[60], JSAX.
7. *Applying GGS grammar rules*
8. *Quering DEX-online and OpenAPI.ro.*
9. *Applying a filter to eliminate numbers and data words*

The names associated entities types are evaluated. If one type is assigned, then labelling will be done immediately. Otherwise, the classification process continues, the main purpose being now the disambiguation. A heuristic based on some frequency lists is applied here, which acts as a short-term memory, a window of three sentences. More details on this will be provided later. If the state of ambiguity is not resolved by this heuristic, then the word will be labelled as ambiguous, with all possible tags.

A query is addressed to DEX-online with each word of the sentence[61]. If we get a positive answer, then the word is labelled with the corresponding lemma. If a negative answer is obtained, an `EXTRA` attribute with the value `NotInDict` is added.

Table 7.1 and Figure 7.3 show a sample of statistics for one (randomly chosen) period. Appendix 1 contains data for the entire corpus.

Table 7.1 Statistics of total words, including those marked *NotInDict*, for the period 1931-1947

Name	Region	Year	Total NotInDict	Unique NotInDict	Total words	% (Total NotInDict /Total words)
Basarabia Chișinăului	Bessarabia	1931	25	21	204	12.25

[60] JDOM is an alternative, augmented, version of the traditional DOM and SAX that deals with reading and writing XML documents, while providing support for XPaths and XSLs. It integrates DOM and SAX, but uses external parser to create documents. Due to the facilities offered and the needs of the present project, JDOM choice was the best solution.
[61] dexonline.ro

Dimineața	Wallachia	1931	2	2	132	1.51
Aradul	Transylvania	1931	29	21	304	9.53
Basarabia	Bessarabia	1932	77	58	1710	4.5
Curierul	Bessarabia	1932	53	44	1676	3.16
Sfatul țării	Bessarabia	1932	20	18	292	6.84
Curentul	Wallachia	1932	17	15	681	2.49
Solidaritatea	Bessarabia	1932	7	7	200	3.5
Luminătorul	Bessarabia	1933	18	16	525	3.42
Adevărul	Wallachia	1933	10	6	100	10
Curentul	Wallachia	1933	2	2	77	2.59
Universul	Wallachia	1933	6	6	199	3.01
Curierul creștin	Transylvania	1933	26	19	411	6.32
Libertatea. Organ politic muncitoresc	Moldavia	1933	28	17	439	6.37
Luminătorul	Bessarabia	1934	14	13	672	2.08
Universul	Wallachia	1934	7	5	149	4.69
Curierul creștin	Transylvania	1934	25	21	442	5.65
Contabilul	Moldavia	1934	9	7	261	3.44
Solidaritatea	Bessarabia	1934	13	13	279	4.65
Luminătorul	Bessarabia	1935	13	11	396	3.28
Adeverul	Wallachia	1935	80	50	1680	4.76
Curentul	Wallachia	1935	5	4	88	5.68
Familia	Transylvania	1935	95	85	533	17.82
Luminătorul	Bessarabia	1936	80	65	2460	3.25
Universul	Wallachia	1936	21	18	490	4.28
Adevarul literar și artistic	Wallachia	1937	5	3	163	3.06
Curentul	Wallachia	1937	19	14	481	3.95
Basarabia	Bessarabia	1938	9	8	97	9.27
Curentul	Wallachia	1938	2	2	85	2.35
Universul	Wallachia	1938	84	61	1803	4.65
Luminătorul	Bessarabia	1939	14	10	703	1.99
Curierul	Wallachia	1939	39	29	864	4.51
Deșteptarea	Moldavia	1940	6	5	87	6.89
Curentul	Wallachia	1940	10	8	706	1.41
Solidaritatea	Bessarabia	1940	19	18	791	2.4
Basarabia	Bessarabia	1941	38	32	1335	2.84
Luminătorul	Bessarabia	1941	13	12	568	2.28

Buletinul Arhiepiscopiei Chișinăului	Bessarabia	1942	9	9	326	2.76
Viata economică din Bălți	Bessarabia	1943	8	7	275	2.9
Buletinul Arhiepiscopiei Chișinăului	Bessarabia	1944	1	1	73	1.36
Discurs Onufrie Andrus RSSM	Bessarabia	1945	3	2	46	6.52
Ethos	Bessarabia	1945	5	5	256	1.95
Scânteia	Wallachia	1947	1	1	153	0.65

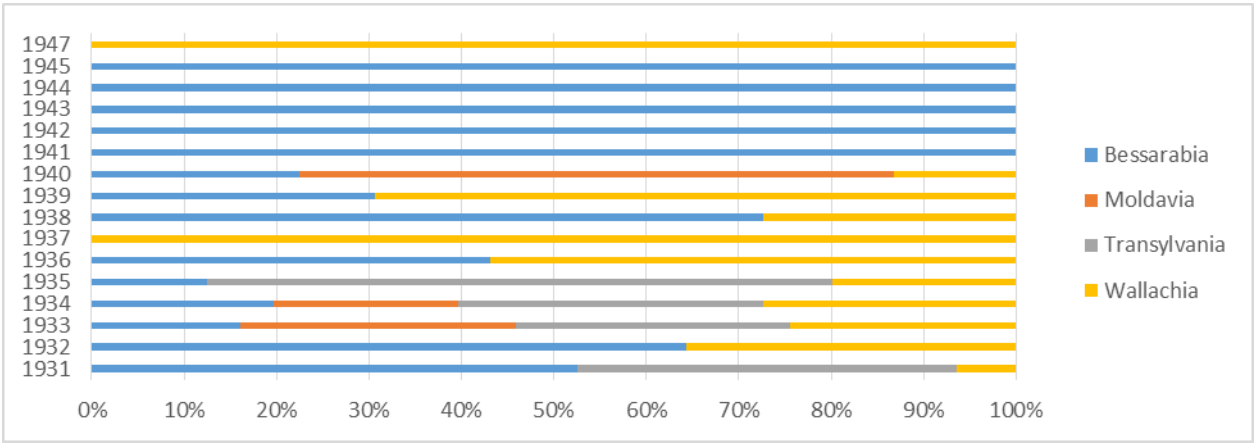


Figure 7.4 The percentage of NotInDict for each region (1931-1947)

For the period 1931-1947, selected as a sample, the corpus from Bessarabia (total words - 12884, total NotInDict – 439 and unique NotInDict – 370) was bigger that from Wallachia (total words - 7851, total NotInDict – 43 and unique NotInDict – 29), from Transylvania (total words - 1690, total NotInDict – 175 and unique NotInDict – 146), and from Moldavia (total words - 787, total NotInDict – 43 and unique NotInDict – 29). However, the situation changes as percentage of Total NotInDict /Total words as follows: Transylvania (9,83%), Moldavia (5.57%), Bessarabia (4.06%), and Wallachia (3,72%), which shows that the newspapers language from Transylvania was visible different from today.

7.1.2 Tracking chronology of old Romanian words and recovering old toponyms

This section proposes a method of chronologically recovering the lemmas of words tagged as `NotInDict` (output from *Text Processing*), using eDTLR. Moreover, this list with unidentified words in DEX-online constitutes a preliminary stage for building a lexicon of old forms (words with significant differences of morphology and syntax).

Furthermore, we are also interested to create a gazetteer with old Romanian toponyms in order to extend the NER (*Name Entity Recognizer*):

For instance: `Valahia`, old name of the county of Wallachia.
`Apulum`, old name of the city of Alba-Iulia.

Each `NotInDict` is marked with two labels (year and region), which attests their origin. Looking in eDTLR for each old entry, we can improve the chronological data (by taking into account the year(s) of the quotation where the `NotInDict` is found).

Moreover, from this list of unidentified words (initially counting 61.137 entries) we have kept only the eDTLR citations whose chronology is precursory of the publication's year from where the word has been extracted (Appendix 2). The folowing diagram (see Figure 7.5) displays the steps used for Extracting Quotations:

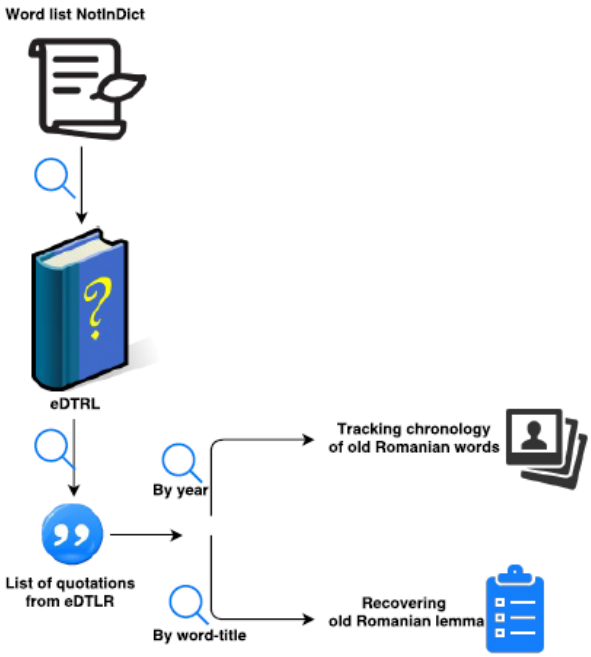


Figure 7.5 The architecture for Exploring eDTRL Quotations

Methodologically, it proceeds as follows:

1. *Collecting all words marked with NotInDict*.

Each `NotInDict`, labelled with publication's year and region to which it belongs, is searched in the citations from eDTLR[62]. There are two different ways to access this dictionary:

- *by using the interface*: the online resource is available, accesed by users through an account. The words are indexed by their first letter and they are gouped like in dictionary. When looking for a word, the user receives a list of senses with associated examples. There are also several filters integrated in the interface, which help to constrain the search.

- *by accessing the inner source*: these are two XML files that were used in the back-end of this online application: eDTLR itself and eDTLR's chronology. The first document includes all entries with the associated definitions and their lists of quotations. Each quotation includes the name of the author and some details regarding the opera.

Here follows an example from DA (Dictionary of the Academy – the first series of volumes in DTLR):

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<entry accente="2°." diacritice="2°.">
    <list>AVAÉT 2°. |   AVAÍU,-ÍE   n-4</list>
    <sense class="0" value="AVAÉT">
        <definition>
            <MorfDef>s. a. </MorfDef>
            <ItalMarker>Taxe, rente, redevance
                <norm> (perçue comme revenu au profit de certains
fonctionnaires de l'État).</norm> Part convenue qu'on rend au
propriétaire d'une brûlerie sur l'eau-de-vie de prunes qu'on
<norm> fait fabriquer chez lui. (În epoca fanariotă) Dare, impozit sau
taxă ce se plăteà pentru diferite beneficii, slujbe, numiri în slujbă,
drepturi etc. Aceste </norm>avaeturi<norm> erau, pe vremurile acelea
de jaf obștesc, foarte numeroase și dedeau venituri considerabile
slujbașilor mici și mari, chiar și domnitorului și membrilor casei
sale.</norm> Luându-și gelatul avaetul lui, după obiceiu, au dat
trupul de l-au îngropat...
                <norm><SG>E. KOGĂLNICEANU, LET.III,250/
            <sub>3</sub>.</SG></norm>
Avaetul caftanelor ce îmbracă Domnul pe boieri: ele la vel log. până
la vel post. câte 100 <norm> lei, de la vel spat. pană la vel stol.
```

[62] According to the lexicographic norms of the Academy, the Dictionary includes all words registered in documents and texts. It includes etymology, and each word sense is illustrated by quotations from a large collection of texts, attributed to social and cultural domains (2500 titles from approx. 4000 volumes).

câte 80 lei. (a. 1776),
<SG>**ap. HEM. 2140.**</SG></norm>
Venitul lui vel cahvegiu: de toată cahfeneaoa... să iea avaetul său, după tocmeala ce se va puteà așezà.
<norm><SG>**ibid. 2141.**</SG></norm>
Vama românească la Ribnic, la havaetu' Do[a]mnei, și la ca[i]macanul, și la ispravnic, la porc 26 pa[rale].
<norm> **(a. 1792),** <SG>**ap. IORGA, S. D. XII, 115.**</SG>
</norm>
Avaetul horelcii ce vine de peste munți.
<norm> **(Mold., a. 1814).** <SG>**URICARIUL, I, 214/**
_{**1**}.</SG>
</norm>
Să plătească numai 30 parale havaet, de pecetluit.
<norm> **(Mold., a. 1803),** <SG>**ibid. II, 180/**_{**12**}.</SG> **mira107015**
</norm>
Nu sânt supuși... a plăti nimica ocârmuirii, sau havaet, sau zăciuală.
<norm> **(Mold., a. 1827),** <SG>**ibid. II, 208/**
_{**32**}.</SG>
</norm>
Să ieie avaetul vătăjiei de arabagii, din zece lei unul.
<norm> **(Mold., 1815),** <SG>**ibid. IV, 206/**
_{**3**}.</SG>
</norm>
Avaetul acestor școli.
<norm> **.. (Mold., 1827),** <SG>**ibid. VII, 92/**
_{**8**}.</SG>
</norm>
Două mii [de] lei am dat havaetul mezatului.
<norm>
<SG>**FILIMON, C. II, 614.** </SG>
</norm>
</ItalMarker>
</definition>
<sense class="**12**" value="**2°.**">
<definition>
<SpSpecDef> **S p e c.** </SpSpecDef>
<SpecDef>**(Astăzi, în Muscel)**</SpecDef>
<ItalMarker>**Avaiet**
<norm> </norm>
</ItalMarker>
<RegDef>**= „zeciuală ce se ia celor ce fabrică țuică la povernele altora; plata fabricatului".**
<SG>**RĂDULESCU-CODIN.**</SG>
[În sensul adesea
</RegDef>
</definition>
<sense class="**20**" value="**|**">
<definition>
<RegDef> **Plur.** **-eturi]** **mira103811**</RegDef>
</definition>
</sense>
</sense>
<sense class="**0**" value="**AVAÍU,-ÍE**">

```
        <definition>
            <MorfDef>adj. </MorfDef>
            <RegDef>v. </RegDef>
        </definition>
    </sense>
  </sense>
    <MorfologicalPart>
        <p>- Din turc.<b>
            'avaïd</b> „rente".
            <SG>CIHAC, II, 544.</SG>
        </p>
    </MorfologicalPart>
</entry>
<entry accente="A" diacritice="A">
<list>Á |   n-2</list>
```

2. *Extracting the first occurrences of related* `NotInDict` *words in a text.*

In addition, we will have the ability to return some quotations in the case that we have them in eDTLR. Here, each title word has associated a set of citations extracted from the Romanian literature, having attached the year of publication. eDTLR is represented by two different XML files:

- the first XML file contains some quotations which include the word marked with `NotInDict`.

Here is the dictionary entry for *dezmeţit*, old form for *dezmeticit* → (EN) *sobered*

```
<entry accente="DEZMEŢIT" diacritice="DEZMETIT">
    <list>DEZMEŢÍT, -Ă   n-2</list>
    <sense class="0" value="DEZMEŢÍT, -Ă ">
        <definition>  <MorfDef>adj.
</MorfDef><SpecDef>(Învechit)</SpecDef>
        <RegDef> Dezmeticit (<b>II 1</b>). </RegDef>
        <DefExemList>
            <exemPair  year="1857" year2="1858">
                <DefExem>Trotuşul e vechiul lor prieten, el îi
primeşte în răcoroasele sale ape fără primejdii, îi îmbrăţoşază şi-i
trimite pe acasă scăldaţi şi dezmeţiţi. </DefExem>
                <SG><AUTHCITE author="NEGRUZZI" pages="315"
sigla="NEGRUZZI, S." source="S." volume="I">NEGRUZZI, S. I,
315</AUTHCITE>, cf. <AUTHCITE author="PONTBRIANT" sigla="PONTBRIANT,
D." source="D.">PONTBRIANT, D.</AUTHCITE>, LM.  </SG>
            </exemPair>
            <exemPair  year="1881" year2="1882">
                <DefExem>Dar azi vălul cade, crudo! dizmeţit din
visuri sece, Fruntea mea este trezită de al buzei tale-ngheţ.
</DefExem>
                <SG><AUTHCITE author="EMINESCU" pages="30"
sigla="EMINESCU, O." source="O." volume="I">EMINESCU, O. I,
```

```
30</AUTHCITE>, cf. <SRCCITE source="DDRF">DDRF</SRCCITE>, RESMERIȚĂ,
D.</SG>
                <norm> - Scris și:  </norm>
            </exemPair>desmeț**it**. - Pl.: <error>[i]dezmeț**i**ți, -te. -
[/i]Și:</error>
        </DefExemList>
        ...
    </sense>
</entry>
```

Here we can find a part of the entry for *dezmeț**i**t*, in whose citations (DefExem) a number of related forms can be found (here: *dezmeț**i**ți*, *dizmeț**i**t*). The field that helps us to find a year is the content of the <SG> tag. The information extracted after parsing this file is stored in some customary structures. Then, the year associated with the <SG> tag is found in the chronology. The chronology file lists the eDTLR bibliography: details regarding authors and publications. For instance, the above example includes as value of an <SG> tag: I. NEGRUZZI, S. I, 315. This content is searched for in this file and the value of the attribute an is extracted from there:

```
<RandSigla nr_item="2686">
    <Cronologie an="1889">168</Cronologie>
    <Sigla>I. NEGRUZZI, S. I–VI</Sigla>
    <Autor> Iacob Negruzzi</Autor>
    <Titlu>Scrieri complecte.</Titlu>
    <Detalii> Vol. I–VI. I: Copii de pe natură. Scrisori; II: Poesii;
III: Mihai Vereanu. Primblări prin munţi; IV: Teatru; V: Pe malul
mării. Din Carpaţi. Traduceri din Schiller; VI: Traduceri din
Schiller. Bucureşti, Socec &amp;amp; Comp., 1893–1897.</Detalii>
    <auxiliar> Iacob Negruzzi, Scrieri complecte. Vol. I–VI. I: Copii
de pe natură. Scrisori; II: Poesii; III: Mihai Vereanu. Primblări prin
munţi; IV: Teatru; V: Pe malul mării. Din Carpaţi. Traduceri din
Schiller; VI: Traduceri din Schiller. Bucureşti, Socec &amp;amp; Comp.,
1893–1897.</auxiliar>
    <Pagini>838</Pagini>
    <NrPaginiVolum>I-492, II-346</NrPaginiVolum>
    <locatie>IFRI</locatie>
    <VecheRara/>
    <COPYRIGHT>NU</COPYRIGHT>
    <stiute>1</stiute>
    <copyright>0</copyright>
    <slavona>0</slavona>
    <SCANATE>scanat doar I si II, restul scanat BCU</SCANATE>
    <scanate>1</scanate>
</RandSigla>
```

3. Adding values extracted for each symbol in the custom structure.

We search for words containing at least part of the sigla tag. In this case, the year, because we

notice that the second document is not complete.

4. Monitoring the final list of items.

This list contains: the main word, a list of principals' quotations found for this word and the year of publication (if it is found in the choronology file).

The final output will be a list with the old forms, quotations and associated years. A few examples are included in Table 7.2.

Table 7.2 Recovering of old word forms in the corpus (a sample)

PhD corpus				**eDTLR**		
NotInDict	**Year**	**Region**	**Newspaper**	**Year**	**Quotations**	**Source data**
abie	1868	Transylvania	Federațiunea	1750	Un om… căzut-au în tălhari de-l jăfuire și-l răniră și-l lăsară **abie** viu în cali	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
acestia	1854	Transylvania	Foaia pentru Minte, Anima și Literatura	1704	Vredniciile **acestia** să cuvin lui Petru, de vreme ce el... au întins pururea, ca alt soare, de taină, peste tot pământul razele darurilor și a facerilor de bine.	Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
semni	1950	Transylvania	Vatra românească	1773	Pi **semni**, omule, tu ești turuielic	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).

simtu	1853	Transylvania	Telegrafulu Romanu	1600	Cum tămăduieşti pre alţii, ce **simtu** mai streini, aşea vino de tămăduiaşte şi pre prietinul tău, pre Lazăr	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
sîngur	1950	Transylvania	Vatra românească	1762	Vel-cupar, cu tipsia cea de argint, stînd înaintea domnului, pune vutcă în doao păhare şi întîi cuparul ia credinţă, apoi domnul **sîngur** ia paharul de-l pune la gură şi ia şi confeturi.	Dan Simonescu, Literatura românească de ceremonial. Condica lui Gheorgachi, 1762. Studiu şi text de... Bucureşti, Fundaţia Regală pentru Literatură şi Artă, 1939. [Secţia istorică, III. Studii].
soarile	1848	Bessarabia	Candela	1750	La Eghipet nu cade niciodată zăpadă, pentru că este **soarile** mai aproape şi în veac **soarile** şi văzduhul este mai herbinte, precum vestesc cei ce lăcuiesc acolo	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
svârli	1950	Transylvania	Vatra românească	1688	Şi mâncă Iacov şi să sătură şi **svârli** cel iubit	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).

The information is stored in a database, but the large number of words and the quantity of associated information can slow down the access. The quotations may help to extend our possibilities to get the correct years, but also they give important data in terms of time and space.

Here an example based on eDTLR process:

- **Initial Input – text form**:

PRESSA, BUCURESCI, 3 AUGUST 1878 (which was classified in this period 1840-1917)

[..] *Dar toate aceste modificări de Constituţiune, adică recunoscerea şi aplicarea principielor puse în tractatul din Berlin, se pot face altfel, decât prin o Constituantă! La aceste modeste cuestiuni rugăm să ne răspundă, limpede şi categoric, ghibacii, isteţii şi mult naţionalii redactori de la Românul.*

- **Initial Input – list of words** that were marked `NotInDict`.

[*Constituţiune, tractatul, cuestiuni, ghibacii*]

- **Output:**

- **Search by word with definition and quotations:**

The only word from the previous list that we found in eDTLR is *ghibac* → (EN) skilled.

```
<entry accente="GHIBACIU" diacritice="GHIBACIU">
    <list>GHIBÁCIU, -CE ||  ||   n-4</list>
    <sense class="0" value="GHIBÁCIU, -CE">
    <definition>
        <MorfDef>adj., adv. </MorfDef>
        <ItalMarker>Adroit. Adroitement. -
            <norm>
                Îndemânatec și isteț (la lucru), iscusit, meșter,
abil, cu apucătură. A d j.</norm>Omeșteșugie ghibace, de care ne putem
cu dreptul mirà.<norm>
                <SG>RUSSO, S. 60.</SG>
            </norm>
                Babele... sânt foarte dibace în tragerea bobilor.
            <norm>
                <SG>ALECSANDRI, P. P. 234.</SG>
[ … ]
```

- **Search by word found in other quotations:**

a. *După ce Petru încheie un* ***tractat*** *demărcătoriu ... între China şi Rosia ... socoti a-şi organiza o armie regu-lată.* - SĂULESCU, HR. II, 456/19

b. *Bogdan însă dizvălea o politică ghibace*. - ASACHI, S. L. II, 24

- **Seach by the authors found:**

SĂULESCU, HR. II, 456/19 - 1834

```
<RandSigla nr_item="3124">
    <Cronologie an="1834">93</Cronologie>
    <Sigla>SĂULESCU, GRAM. ROM. I–III</Sigla>
    <Autor>Gh. Săulescu</Autor>
    <Titlu>Gramatică românească sau observaţii grămăticeşti asupra limbei româneşti. Pentru şcoalele normale şi ghimnaziale. Partea I-ia etimologhică, II sintactică, III de prosodia limbei româneşti şi de versificaţie, la care s-au adaos mai multe exempluri de versuri.</Titlu>
    <Detalii>Iaşi, Tipografia Sfintei Mitropolii, partea I–II: 1833; partea a III-a: 1834.</Detalii>
    <auxiliar>Gh. Săulescu, Gramatică românească sau observaţii grămăticeşti asupra limbei româneşti. Pentru şcoalele normale şi ghimnaziale. Partea I-ia etimologhică, II sintactică, III de prosodia limbei româneşti şi de versificaţie, la care s-au adaos mai multe exempluri de versuri. Iaşi, Tipografia Sfintei Mitropolii, partea I–II: 1833; partea a III-a: 1834.</auxiliar>
    <Pagini>234</Pagini>
    <NrPaginiVolum>I – 234</NrPaginiVolum>
    <locatie>IFRI</locatie>
    <VecheRara>CV</VecheRara>
    <COPYRIGHT>NU</COPYRIGHT>
    <stiute>1</stiute>
    <copyright>0</copyright>
    <slavona>1</slavona>
    <SCANATE>scanat</SCANATE>
    <scanate>1</scanate>
</RandSigla>
```

ASACHI, S. L. II, 24 - 1838

```
<RandSigla nr_item="340">
    <Cronologie an="1838" an2="1839">99</Cronologie>
    <Sigla>ASACHI, S. L. I–II</Sigla>
    <Autor>Gh. Asachi</Autor>
    <Titlu>Scrieri literare.</Titlu>
    <Detalii> [Vol. I şi II]. Ediţie îngrijită, cu prefaţă, note şi glosar de N. A. Ursu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, [1957]. (Biblioteca pentru toţi).</Detalii>
    <auxiliar>Gh. Asachi, Scrieri literare. [Vol. I şi II]. Ediţie îngrijită, cu prefaţă, note şi glosar de N. A. Ursu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, [1957]. (Biblioteca pentru toţi).</auxiliar>
    <Pagini>654</Pagini>
    <NrPaginiVolum>I – 286; II – 368</NrPaginiVolum>
    <locatie>IFRI</locatie>
    <VecheRara>–</VecheRara>
    <COPYRIGHT>DA</COPYRIGHT>
    <stiute>1</stiute>
```

```
    <copyright>1</copyright>
    <slavona>0</slavona>
    <SCANATE>scanat</SCANATE>
    <scanate>1</scanate>
</RandSigla>
```

We can say that the database based on tracking chronology of old Romanian words from our journalistic corpus using quotations from eDTRL can serve also to recover their old lemma (more in next section).

7.1.3 Recovering old lemma

To recover old lemma using eDTLR is an objective began a few years ago and described in [Cristea *et al.*, 2012]. By building a list of old lemma we can provide the basis for the implementation of a diachronic POS-tagger.

Based on the word list `NotInDict`, extracted from our historical corpus, every old entry has been searched for the word-title in eDTLR. This resulted in **861** entrances with a correspondent in eDTLR. If we exclude the duplicates (there are old words encountered in several publications of the four regions analyzed), we remain with **582** old lemmas.

For illustration, we mention some examples.

Ex_1: old lema – `desbinare` → present lema – `dezbinare`

Our corpus				eDTLR				
NotInDict	**Year**	**Region**	**Publication**	**Year**	**Quotations**	**Lemma**	**Sigla**	**Title**
desbinare	1949 1893	Transylvania Moldavia	Vatra românească Convorbiri literare	1897	La cas de **desbinare** între tată și mumă, consimțimântul tatălui este de ajuns.	DEZBINARE	HAMANGIU, C. C.	Const. Hamangiu, Codul civil român. Conform textului oficial. Precedat de Constituţiune şi un index alfabetic. Bucureşti, Carol Müller, 1897. (Biblioteca pentru toţi). B. Hasdeu, Istoria critică a românilor. Pământul Țărei Românești. Volumele I–II. Bucureşti, Imprimeria Statului şi Tipografia Antoniu Mănescu, 1875
desbinarea	1890 1919 1949 1877	Transylvania Wallachia	Aradu Patria Vatra românească Romanulu	1872	S-au găsit printre români Oameni ... Carii ... sămănând **desbinarea** pe pământul strămoşesc, Au scăzut ... puterea neamului lor.	DEZBINARE	ALECSANDRI, T.	Vasile Alecsandri, Opere complete. Partea întâia: Teatru. Volumul I: Canţonete comice, scenete şi operete; volumul II: Vodeviluri; volumul III: Comedii; volumul IV: Drame. Bucureşti, Editura Librăriei Socec & Comp., 1875.
desbinări	1878 1877	Wallachia	Romania libera Romanulu	1962	Și pe dânsa-ncăleca. Trei **desbinări** că mi-i da, Pe Iovița mi-l zărea.	DEZBINARE		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.

desbinării	1950	Transylvania	Vatra românească	1868	Nenorocirile politice ce-l băntuiau din causa **desbinării** diferitelor populațiuni.	DEZBINARE	CONV. LIT.	Convorbiri literare. Iași, apoi București. Anul I (1867–1868) ș. u.

Ex_2: old lema – `desbina` → present lema – `dezbina`

Our corpus				**eDTLR**				
NotInDict	**Year**	**Region**	**Publication**	**Year**	**Quotations**	**Lemma**	**Sigla**	**Title**
desbină	1939 1949 1950	Bessarabia Transylvania	Luminătorul Vatra românească	1885	Strâgile ... sărăcesc pe oameni, îi **desbină**.	DEZBINA	SBIERA, P.	Ion al lui G. Sbiera, Poveşti poporale româneşti. Din popor luate şi poporului date de... Cernăuţi, Editura Proprie, Tipografia Arhiepiscopală, 1886.

In Appendix 3, one can see more situations. The main problem which arises is that each `NotInDict` is looked only in quotations. One quotation may be attached both to a title-word different than the word searched for or belonging to the searched for title-word.

7.2 Language variation over time using the MaxEnt model

This section describes a new method for tracing language variation, based on the MaxEnt (*Maximum Entropy*) text classifier being commonly used in Natural Language Processing (NLP) tasks, first introduced by Berger *et al.* [1996] and Della Pietra *et al.* [1997] in statistical estimation and pattern recognition. Noteworthy is that the MaxEnt classifier has great results when the training corpus is limited, as is our case. Furthermore, this study is continuing the work presented in [Gîfu & Simionescu, 2016]. Here, we gave up at the prior periodization and we analyzed the journalistic language for each of the four geographical areas.

Actually, we are aiming to analyze here using this model, from the perspective of transformation patterns and in a supervised manner, the differences between the „source”

lexicons (Moldavian, Transylvanian, Wallachian, and Bessarabian) and the „destination" lexicon (DEX-online). All the substring replacement operations (referred to as "REPs") are classified and extracted from the mentioned correlations list (Appendix 4). Each replacement rule mentions a tripple: the character offset in the source word where the replacement occurs (starting from 0), the initial string and the replacement string. Based on these REPs, a set of fictive/candidate words are generated, each having a trust score attached. If a candidate word is found in the destination lexicon, the two words are marked as a corresponding pair.

Basically, all unknown words (`NotInDict`) are extracted in order to find for them the current correspondent. By applying these operations on the first word of the pair, the second one is obtained.

Example 1: the noun `educației` → (En): `education`, in the oblique case

`educatiunei` - `educației`: {REP(5, t, ț), REP(7, un,)}

where `educatiunei` is the "source" word, and `educației` is the "destination" word.

To avoid excessive segmentation of the replacement operations and to model transformation rules based on longer patterns, we impose a minimum distance of two characters between two REPs operations.

By applying this constraint on the previous example, only one operation is extracted:

`educatiunei - educației: {REP(5,tiun,ți)}`

Example 2: `imperatesci` → (En): `kingly/royal`

The model indicates the following transformations that could be applied at different positions: `REP(0,i,î), REP(3,e,ă), REP(5,a,ă), REP(8,sc,șt)`

A trust score threshold has been used at this point to filter REPs with scores, which have empirically proved of no use. To generate candidate destination words, all possible combinations of REPs are applied in turn on the source/unknown word.

In this example, the candidate words generated by applying all the indicated REPs gives the following list of generated destination candidates:

imperateşti, imperătesci, impăratesci, împeratesci, imperăteşti, impărateşti, împerateşti, impărătesci, împerătesci, împăratesci, impărăteşti, împerăteşti, împărateşti, împărătesci, împărăteşti

By looking up candidate forms[63] in the destination lexicon, which should be a comprehensive resource, word form associations can be inferred automatically. In this case: imperatesci - împărăteşti

Example 3: cetatianii → (En): *citizens*

REPs indicated by the trained model: REP(3, atia, ăţe), REP(4, t, ţ)

The two REPs overlap, so, they cannot be applied synchronously. The current form of the word is obtained by applying the first REP only. The candidate destination forms are:

cetaţianii, cetăţenii

By looking up candidates forms in the destination lexicon, the final word form associated is: cetăţenii.

7.2.1 The corpus

All data for this study is revealed in Table 7.3 and Figures 7.6 and 7.7.

Table 7.3. General corpus statistics

Region	Time	Total considered words[64]	Total unknown words	Total unique unknown words	%(total unknown words/total words)
Moldavia	1829-2015	65901	5085	2979	7.72

[63] When more candidate forms are found in the destination lexicon, the one with the highest trust score is considered.

[64] From the total tokens the punctuation, the numbers and the words with less than two characters were removed.

Wallachia	1829-2015	137261	6525	4105	4.75
Transylvania	1837-2015	160923	21023	8518	13.06
Bessarabia	1817-2015	107324	4703	2891	4.38

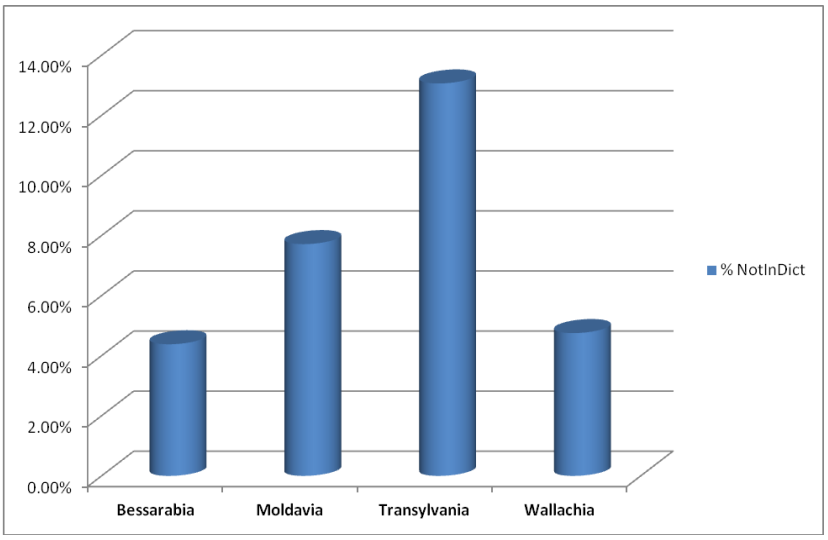


Figure 7.6 Percentages of NotInDict Words - 1817-2015

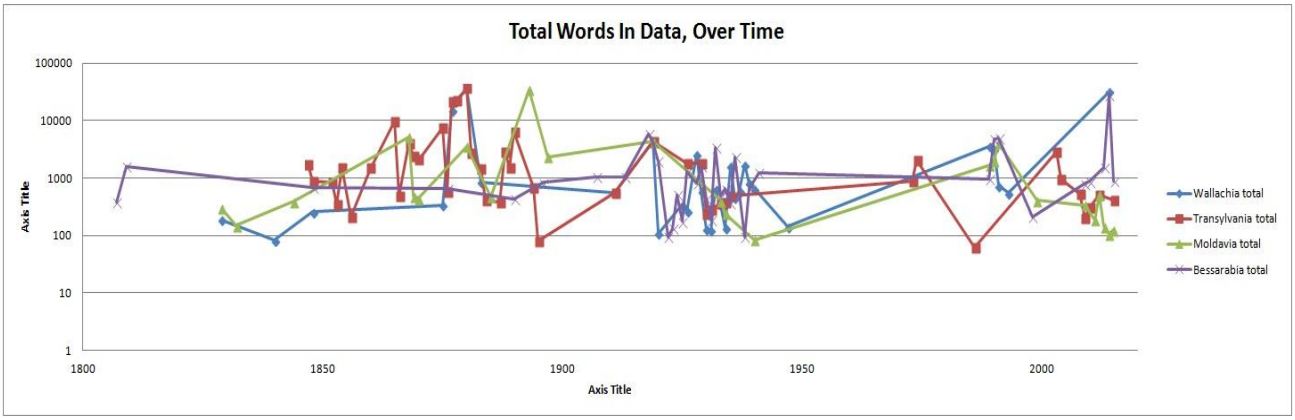


Figure 7.7 The percentage of total words in data over time for each historical part of Romania

In the period 1817-2015 chosen for this study, the corpus from Transylvania (total words - 160923, total NotInDict – 21023 and unique NotInDict – 8518), was bigger that from Wallachia (total words - 137261, total NotInDict – 6525 and unique NotInDict – 4105), from

Bessarabia (total words - 107324, total NotInDict – 4703 and unique NotInDict – 2891), and from Moldavia (total words - 65901, total NotInDict – 5085 and unique NotInDict – 2979). Furthermore, we observe that the percentage of Total NotInDict /Total for Transylvania (13,06%) is also higher than in the other three analyzed regions: Moldavia (7,72%), Wallachia (4,75%), and Bessarabia (4,38%). All these results show that the newspapers language from Transylvania was different from the other regions of Romania, since the first newspapers in Romanian appear in a small number, a decade later.

7.2.2. Eperiments and results

Using the mechanism of automatic correlation of unknown words (marked with NotInDict), various experiments were conducted on the available corpora. An evaluation of the model is presented in Table 7.4. The training data for each period/region consists of a correlation list, which associates approx. 50% of the old words from the corpus with their forms in contemporary language use. The precision for various training scenarios is presented. Given the relatively limited data, a 10 fold cross validation was used. The value indicated for each dataset represents the successful automatic correlation percent of the evaluation data.

Table 7.4 Evaluation of the automatic correlation of unknown distinct words mechanism.

Wallachia 1829 - 2015	**Transylvania 1838 - 2015**	**Moldavia 1829 - 2015**	**Bessarabia 1817 - 2015**	**All texts**
82.93%	81.40%	87.06%	86.82%	84.55%

At the evaluating timing, we talk about the differences in data volume that builds up and with the noise between regions. The cumulative accuracy per total is over 84%.

Table 7.5 presents various detailed aspects of the data and the percent of correlations that were automatically performed by the presented mechanism.

Table 7.5 Unknown words in corpus and their automatic correlations coverage.

Region / Time	**Wallachia 1829 - 2015**	**Transylvania 1838 - 2015**	**Moldavia 1829 - 2015**	**Bessarabia 1817 - 2015**

Percent of known words (%)	94.98	85.96	91.82	95.91
Total distinct correlations	2013	6870	1628	970
Total occurrences of correlations	4066	17132	2939	1429
Identifiable words (%)	98.11	97.40	96.55	97.91

The percent of correlated words decreases in more recent periods of time, while the number of known words percent increases. The closer the period of time is, the more words are found in the contemporary morphologic dictionary. The decreased correlation percent in the more recent texts is explained by the fact that the remaining unknown words are mostly foreign words, or proper names, while the quantity of words deformed based on some observable lexical evolution phenomenon is reduced.

In table 7.6 (viewing data in Figure 7.8) shows the most frequent REPs (12) operations extracted from texts, clustered of different regions since the second decade of the 19th century.

Table 7.6 Example with frequent REPs distribution per time frame and region

REPs	**Wallachia 1829 - 2015**	**Transylvania 1838 - 2015**	**Moldavia 1829 - 2015**	**Bessarabia 1817 - 2015**
u ->	2.61%	18.26%	11.55%	5.79%
s -> z	12.89%	5.19%	9.06%	4.06%
e -> ă	5.91%	3.83%	6.22%	1.66%
ei -> ii	6.27%	2.56%	3.58%	7.86%
e -> i	2.61%	2.10%	3.09%	2.07%
i -> î	0.75%	1.99%	5.82%	1.82%
i -> e	3.22%	1.23%	3.04%	3.23%
ĭ -> i	2.04%	1.77%	1.14%	1.99%
a -> ă	1.22%	1.97%	1.19%	1.16%
s -> x	3.58%	1.48%	0.70%	0.08%

a ->	1.97%	1.30%	2.39%	2.73%
e -> î	1.72%	1.38%	2.34%	0.91%

For instance, the REP "u ->" deletes a character from words, usually in the suffix. A synthesis about the rules that have increased the phonological character of the Romanian orthography we can find in [Hristea, 1981]. In our case, this REP represents almost 30% of all the REPs founded in the Romanian texts from 1829-1917, but it usually appears at the end of words.

Today, the character "u" at the end of the words disappeared from words or grammatical forms which are ending in:

- *diphthongs* as: *ai, îi, ei, oi, ui*
- *consonants* [k'], [g'] and [č]followed by *i* non-syllabic.

Ex: *luai* → (En) take *lămâi* → (En) lemon *mei* → (En) my
auzii → (En) hear *pietroi* → (En) ock *cui* → (En) nail
ochi → (En) eye *vechi* → (En) old *unghi* → (En) angle...

This type of words are written with "u" at the end, although it had vanished long time ago from pronunciation.

Another situation is when "u" at the end of the words appears in the current spelling only in the contexts that play a phonetic reality and must be pronounced:

ambulatoriu → (En) ambulatory *beneficiu* → (En) benefit
contradictoriu → (En) contradictory *interogatoriu* → (En) interrogation
obligatoriu → (En) obligatorily *onorariu* → (En) honorarium
salariu → (En) salary *serviciu* → (En) service
teritoriu → (En) territory…

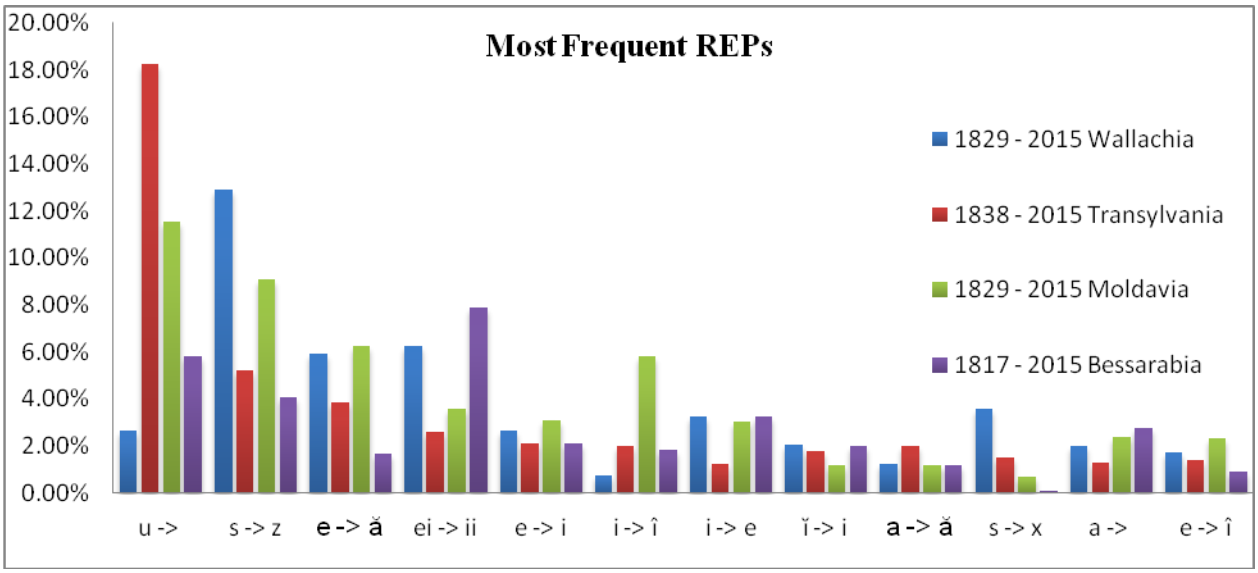


Figure 7.8 The distributions of the 12 most frequent REPs per time frame and region

Below, in Table 7.7 for each of the most common REPs one example is given.

Table 7.7 Example with frequent REPs distribution per time frame and region

REPs	**Romania (old word form – present word form)**		**English term**	**Republic of Moldavia (old word form – present word form)**		**English term**
u ->	introdusu	introdus	*introduced*	comisu	comis	*commit*
s -> z	sburatu	zburat	*flown*	desvoltarea	dezvoltarea	*development*
e -> ă	doue	două	*two*	semenat	semănat	*seeding*
ei -> ii	apunerei	apunerii	*go down*	inimei	inimii	*heart*
e -> i	stenge	stinge	*slake*	aceeași	aceiași	*same*
i -> î	imperatu	împeratu	*emperor*	intre	între	*between*
i -> e	cari	care	*which*	aducire	aducere	*bringing*
ĭ -> i	tocmaĭ	tocmai	*just*	observărĭ	observări	*observations*
a -> ă	lasați	lăsați	*leave*	învața	învață	*learn*
s -> x	esecutată	executată	*executed*	esistă	există	*exist*
a ->	salutea	salute	*salute*	categoriea	categorie	*category*
e -> î	cuventul	cuvîntul	*word*	penze	pînze	*canvas*

One can observe that some REPs operations present a higher than average frequency for both analyzed regions, in the same period of time, like the first and the third REPs, for the first

period and the second period, respectively.

Of course, these are the simplest situations, when it is just one REP operation. However, in the present corpus, we have complex cases, when several REPs operations have intervened. To increase the accuracy of statistical data in identifying automatically the correlations, we decided to focusonly to three REPs operations for each words pair (old - new).

To illustrate this option, here is one example for each geographical area that includes three REPs operations. More examples in Appendix 4:

Transylvania: a noun `serbâtoria` → (En): `celebration` in the direct case:

```
serbâtoria = sărbătoarea: î->ă ria->area e->ă
```

where: `serbâtoria` is the "source" word, and `sărbătoarea` is the "destination" word.

Wallachia: a noun `esposiţiunea` → (En): `exhibition` in the direct case:

```
esposiţiunea – expoziţia: s->z unea->a s->x
```

Moldavia: a predicative verb `măngăemu` → (En): `cosset` in the indicative mode:

```
măngăemu = mîngîiem: ă->îi ă->î u->
```

Bessarabia: a noun `iantămplari` → (En): `events` in the direct case:

```
iantămplari = întîmplări: ia->î ă->î a->ă
```

The results from Table 7.8 and charts below are very promising. The "Identifiable words" represents the percent of words from the texts which are known words or which can be correlated automatically with a known form. As we can see, for all geographical area this indicator is over 95%.

Table 7.8 Known words vs. identifiable words comparison

Region/ statistical parameter	Transylvania	Wallachia	Moldavia	Bessarabia
Known words	85.96%	94.98%	91.82%	95.91%
Identifiable words	97.40%	98.11%	96.55%	97.80%
MaxEnt model - Precision	81,39	82,92	77,05	86,81

Moreover, over 96% of the words can be recovered, but only 76% of them are automatically correlated with a precision of 82%. Regarding the indicator "Known Words", only for Transylvania the result is smaller, because the corpus was bigger. In some cases, to identify the current form of the word, three REPs operations were not enough.

Below, in the following charts, we can visualize time evolution of three metrics, for each of the four analysed regions: the percentage of unknown words over time, the percentage of correlations that has been successfully made from amongst of the unknown words, and the percentage of known words + those recovered [65] (based on automatic correlation).

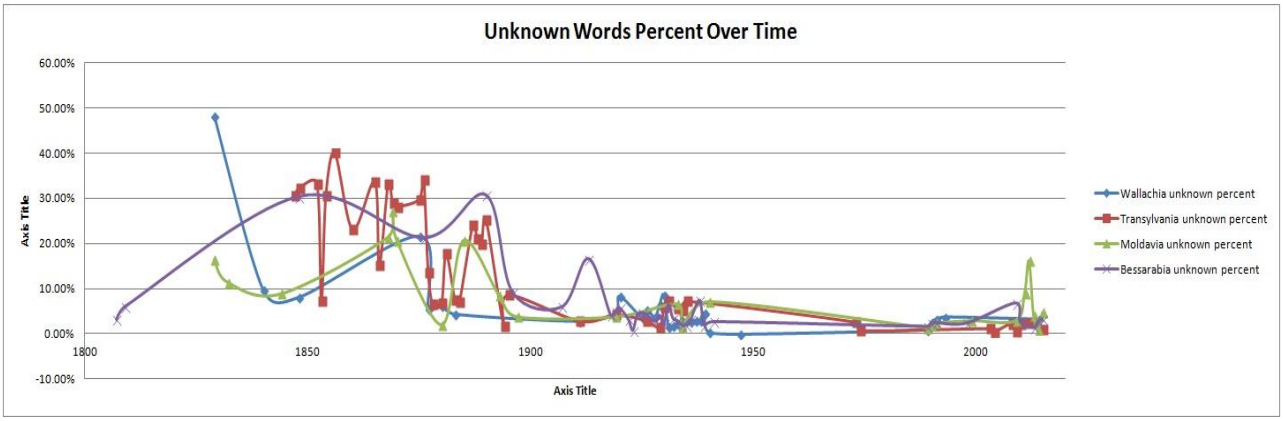


Figure 7.9 The percentage of unknown words over time for each historical part of Romania

Therefore, we find that only from the 20th century those four Romanian idioms begin to overlap, with small deviations in Bessarabia before the outbreak of First World War. In addition, the newspaper language from Bessarabia goes through – as we described already in Chapter 6 – more stages depending on a number of historical factors described previously.

Furtheremore, an interesting case is the newspapers language of Transylvania that is distinguished by a number of particularities in the 18th century. Starting with 1829, when the first

[65] Is a useful metric for morphological disambiguation. Basically, it greatly reduces the POS-tags list to identify the best correlations for an unknown word. In fact, it tells us for what percentage of unknown words we can attach a list of tags.

newspapers appear in Wallachia and Moldova, the two idioms define what we can call the Romanian language in the printing press.

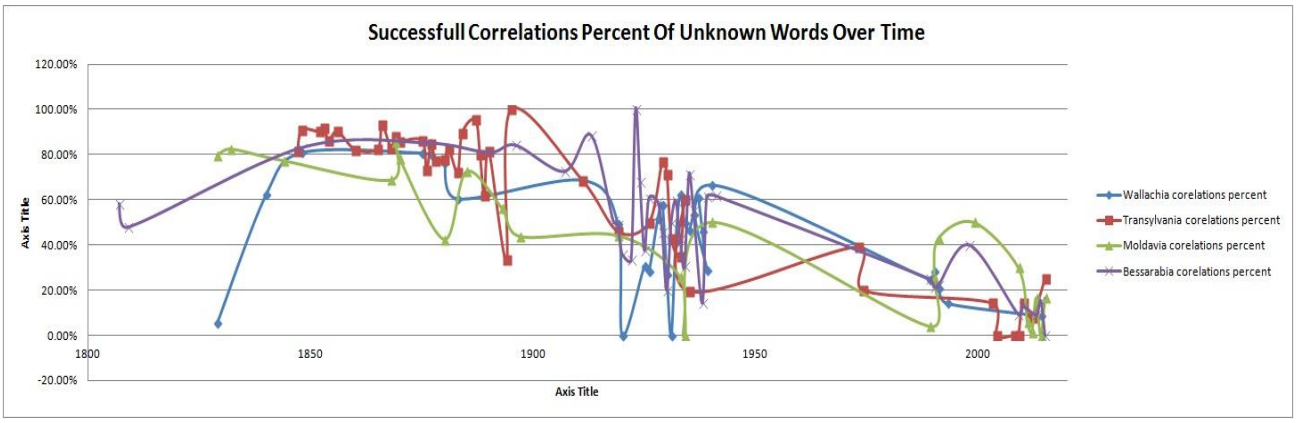


Figure 7.10 The percentage of successful correlations of unknown words over time for each historical part of Romania

In Figure 7.10, we observe that this model, successfully automatically identifies the correlations for unknown words. Each point on the graphic represents a text file. The differences are caused by several factors: amount of words from files, the author's writing style, etc. We notice that in Wallachia, before 1850, the percentage of the automatically identified correlations is extremely low, a situation which is adjusted at 1931/'34. This situation coincides for Moldova as well. The low values for the period after 1989 have as justification the fact that those unknown words are either person's names or abbreviations of parties, organizations, etc., for which the correlations were not found.

For example: IPJ Vrancea, Ion Balerca, SIF, CJ, PNL, PSD, etc.

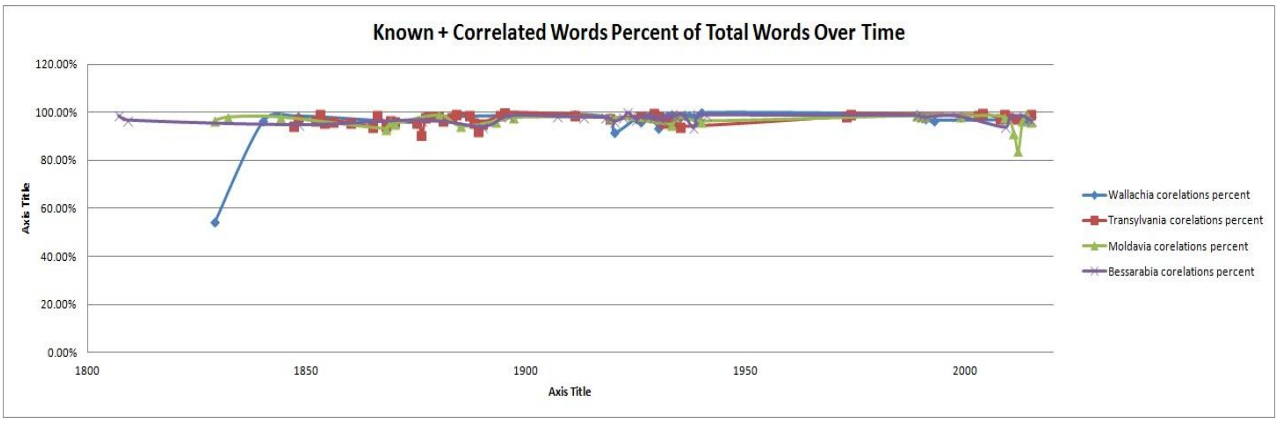


Figure 7.11 The percentage of known words and correlated words of total words over time for each historical part of Romania

In Figure 7.11, the situation is much better, with the exception of the year 1829, when *Curierul românesc* was first published in Bucharest. We note that the last metric can support the implementation of a diachronic POS-tagger based on the POS-tags list automatically identified for unknown words in the current corpus.

Part of the motivation of this research is related to the development of a diachronic POS tagger for Romanian. The MaxEnt model could be used to add enhanced support for unknown words, for a POS tagger for contemporary Romanian. Syntax and orthography play a very important role in POS tagging, and Gold corpora is required for training such a model. However, since this is not available, being able to associate unknown words with their known forms would still be highly beneficial to POS tagging. This would reduce the ambiguity of unknown words in the process of POS tagging diachronic texts.

7.3 Language Similarity Evaluation using SVM and LSTM models

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.

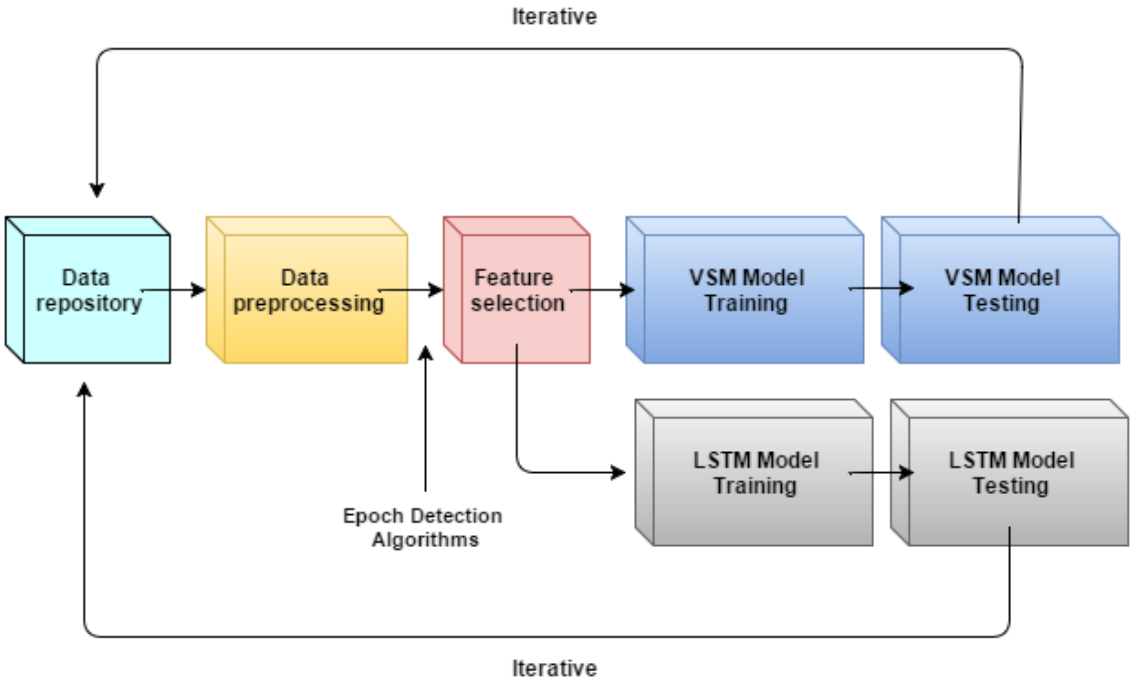


Figure 7.12 Architecture for SVM-LSTM-based epoch detection system

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 (Figure 7.12) on news corpora, which determines the degree of similarity between cognate languages.

The methodology we present is language independent and it 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.

7.3.1 The corpus

A large corpus (over 3 millions lexical tokens and 6550 pages), chronologically ordered, since the second decade of the 19th century, was developed, structured in four independent collections of publications corresponding to Moldavia – 68553 words, Wallachia – 148255, Transylvania – 2294312 words, and Bessarabia – 117063 words described in Chapter 6 (see Appendix 1). Note, that this corpus includes also the Bessarabian texts transliterated from Cyrillic to Latin, given that for the period 1944-1989, the Bessarabian representativeness corpus was basically null, and a few texts from Romania of the years 2000s.

7.3.2 Experiments and results

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 this:

```
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, for the word “pace” (“peace” in Romanian); the word and the occurrences of this word precede 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 (Table 7.9) is determined by a certain distribution of words.

A pair (Bessarabia-Romania) from a typical vectors looks like this:

```
<pair>
w=pace
b_MD=14 pace   2 1817    2 1918    1 1919    2     1920
     2 1929    1 1990    1 1998    1 2009    2 2014
b_RO=768 pace  1 1865    1 1868    17 1877   15    1878
     3 1880    1 1897    2 1919    1 1928    230   1949
     493 1950  3 1989    1 1990
</pair>
```

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

We respectively test whether two samples come from the same statistical population or no. We are also interested to see whether there is a large variance with respect to the mean or no, or if the ratio of change from year to year shows an upward or a downward trend.

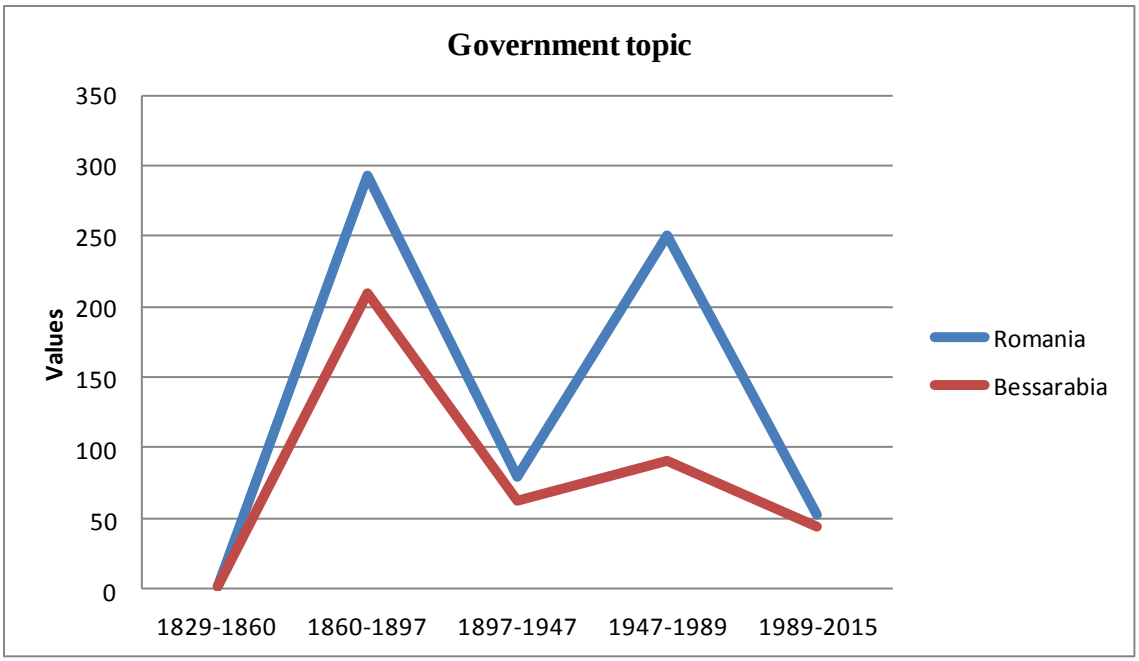
For a very large corpus, like Google books for example, one can chose an arbitrary set of topics to investigate, but in this case we have a limited amount of data. Thus, we need first to indentify the topics that are represented in our corpus. For this, we apply the LDA (*Latent Dirichlet Allocation*) algorithm. At this step, the non photo-lexical vectors are used.

We filtered out set 10 topics (Table 7.9) the following topics for both source corpus and target corpus like:

Table 7.9 Topics statistics

Topics (RO)	Topics (EN)	Epochs									
		Romania					Bessarabia				
		1829-1860	1860-1897	1897-1947	1947-1989	1989-2015	1829-1860	1860-1897	1897-1947	1947-1989	1989-2015
alegeri	elections	0	67	21	141	45	0	1	15	2	47
drept	justice	15	243	30	2003	51	17	210	68	1600	62
economie	economy	7	39	11	357	30	9	41	12	18	37
guvern	government	2	293	80	250	52	2	210	62	90	44
literatura	literature	1	3	30	1516	2	2	2	1	33	9
muncitor	working	0	3	6	1734	3	0	1	5	346	4
pace	peace	2	57	3	872	1	4	34	8	561	5
politic	political	7	265	53	821	51	5	189	43	39	45
popor	people	1	145	40	2698	17	1	4	37	2	24
sat	village	1	27	8	1894	3	2	2	12	1	14

For illustration, below two topics *justice* and *government* plotted:



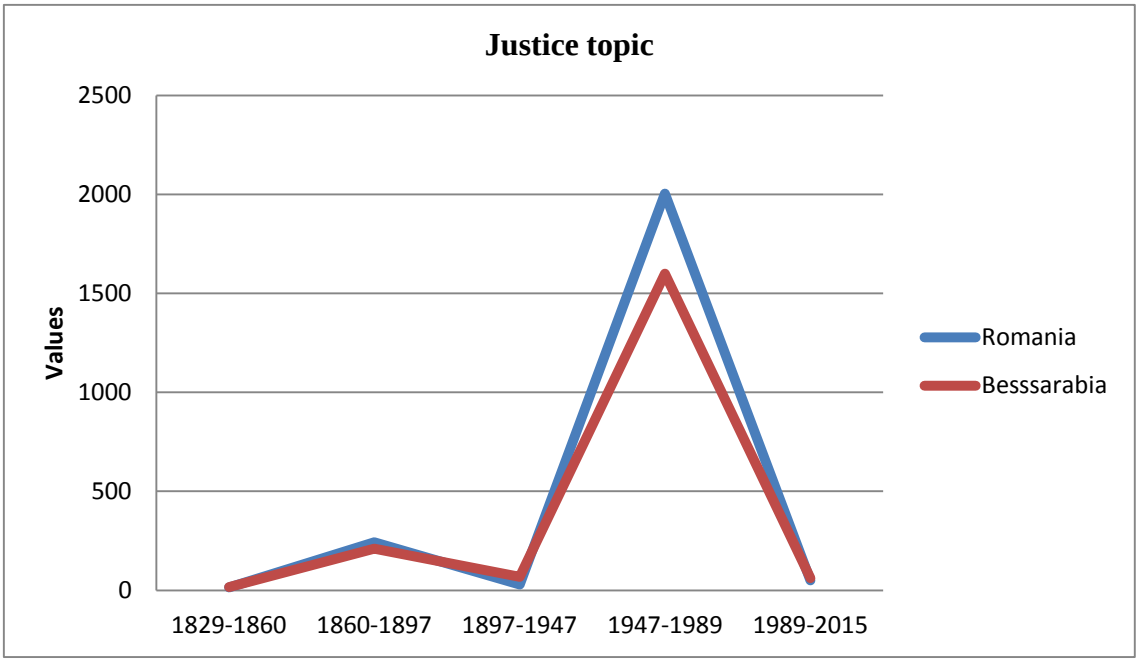


Figure 7.12 Graphical representation for *government* and *justice* over time

For these topics the following epochs have been identified for Romania (in each historical region):

Table 7.10 Romanian Epochs in Newspapers

Transylvania	**Wallachia**	**Moldavia**	**Romania per total**
1847-1860	1829-1865	1829-1844	1829-1860
1865-1895	1875-1911	1868-1897	1860-1897
1919-1950	1919-1947	1919-1940	1897-1947
1973-1986	NA	NA	1947-1989
1989-2015	1989-2015	1989-2015	1989-2015

Considering these epochs as categories we build two classificatory (SVM and LSTM) over whole source corpus (Weka implementation). We classified each news from the target corpus, i.e. Bessarabian. First thing we wanted to check was whether the classification is able to pin point correctly the target 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.

In Table 7.11 and 7.12 we present the figures for each epoch separately, plotted in Figure 7.13 (a, b, c).

Table 7.11 Similarity as SVM classifier confidence

Tranylvania epoch	Bessarabia	Wallachia epoch	Bessarabia	Moldavia epoch	Bessarabia
1847-1860	52%	1829-1865	61%	1829-1844	85%
1865-1895	57%	1875-1911	64%	1868-1897	73%
1919-1950	63%	1919-1947	69%	1919-1940	75%
1973-1986	59%	NA	-	NA	-
1986-2015	73%	1989-2015	84%	1989-2015	82%

Table 7.12 Similarity as LSTM classifier confidence

Tranylvania epoch	Bessarabia	Wallachia epoch	Bessarabia	Moldavia epoch	Bessarabia
1847-1860	59%	1829-1865	62%	1829-1844	89%
1865-1895	63%	1875-1911	68%	1868-1897	75%
1919-1950	68%	1919-1947	70%	1919-1940	79%
1973-1986	62%	NA	-	NA	-
1986-2015	90%	1989-2015	87%	1989-2015	88%

We can see that 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.

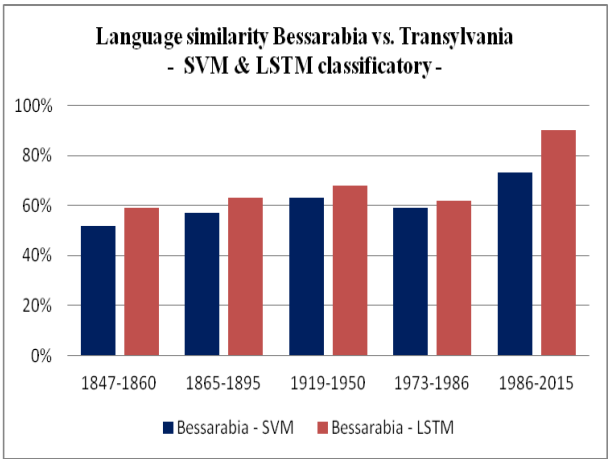
We obtain accuracy (formula 7.1) on epoch prediction of almost 63% (Bessarabia vs. Transylvania), 72% (Bessarabia vs. Wallachia), 81% (Bessarabia vs. Wallachia) with SVM and 70% with LSTM. These results indicate that both classificatory work correctly on the source corpus.

$$Acc = (n / N)*100 \quad (7.1)$$

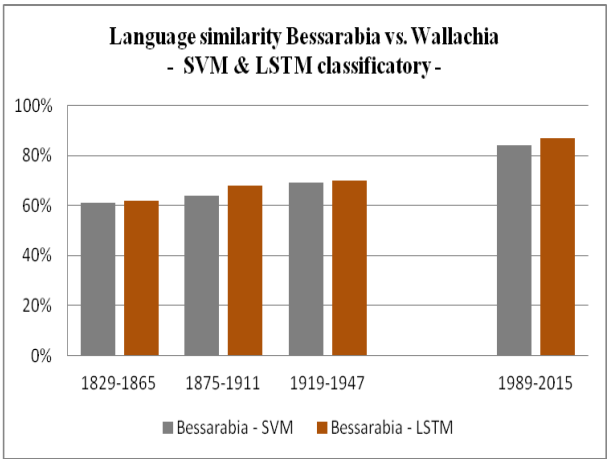
where: n - number of correct classified texts

N - number of the test texts

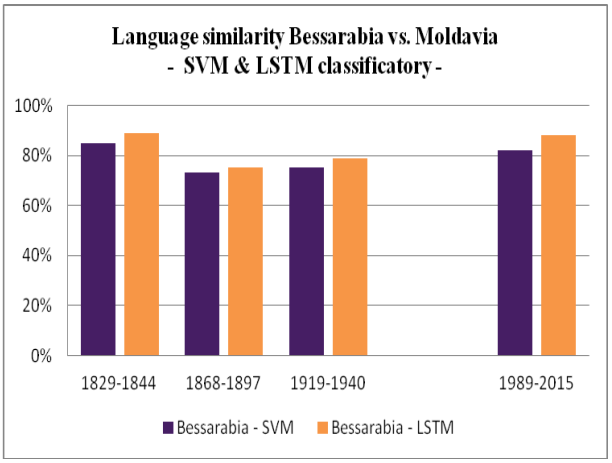
We averaged over the classificatory confidence for each epoch separately. We take this parameter as indicator of the similarities between the cognate languages, because that the classificatory is appropriate, the confidence reflects the similarity.



- a -



- b -



- c -

Figure 7.13 Accuracy comparison between SVM and LSTM for language similarity between Bessarabia & each Romanian region

In Figure 7.14 present a plot that represents the percentage of the Bessarabian regionalisms for each epoch separately. The results confirm that the Tsarist occupation and the "Developed Socialism" epoch significantly have influenced the Bessarabian writing style in newspapers.

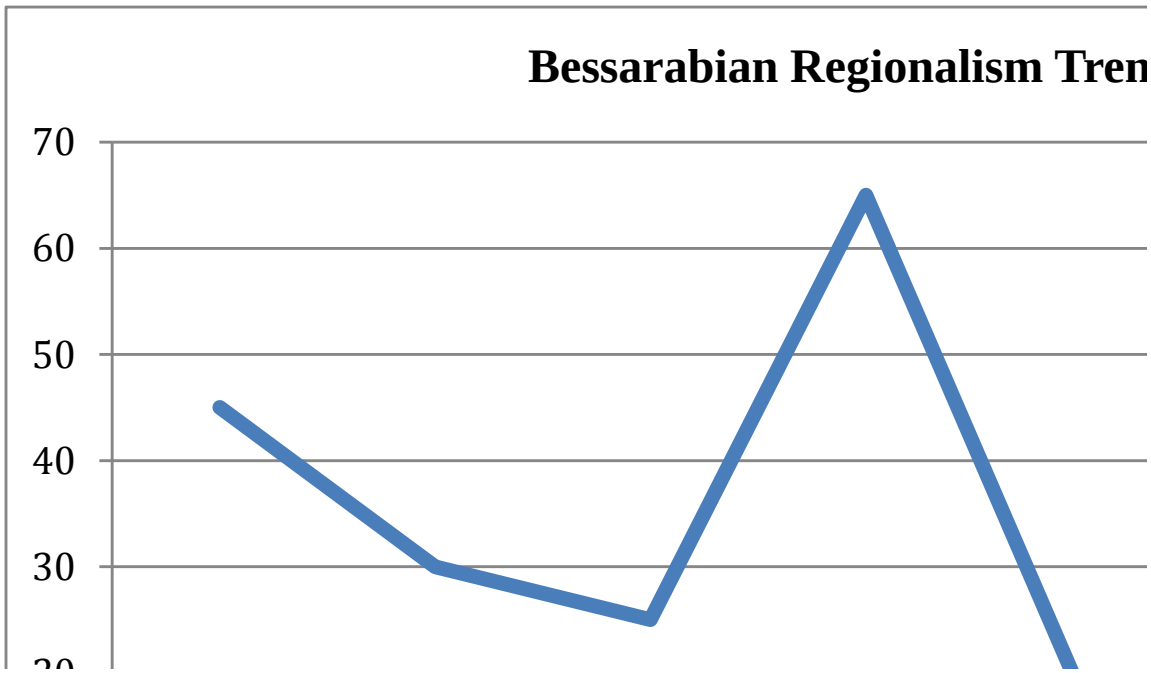


Figure 7.14 Bessarabian regionalism trend over time

7.4 Time Evolution of Writing Styles in Newspapers Language

Below, we present a diachronic pilot analysis centered on the exploration of differences between the writing styles of journalistic texts in Romanian language. This analysis is centered on the time evolution of this language across two adjacent regions, Bessarabia and Romania since 1941. This work was done by using *ReaderBench* framework[66] to calculate a number of textual complexity indices (textual features, textual formality, and textual styles) that can be reliably used to characterize writing style measured in terms of LSA and LDA.

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).

[66] *ReaderBench* is a multilingual platform that integrates a multidimensional model based on Support Vector Machines. [Dascalu et al. 2013]

7.4.1 The corpus

Our corpus was developed based on a newspaper collection and contains around 62,000 lexical tokens (see Table 7.11 for descriptive statistics). Importantly, this corpus represents a first iteration towards building a Romanian Gold corpus centered on diachronic meta-annotation. Note, in this case, the Bessarabian publications written in Cyrillic during 1944-1989 were not included.

Table 7.10 Different SNA metrics applied on both models.

Region	Period	Total words	Sources
Romania	1941-1991	25,873	*Deșteptarea; România literară; Scânteia; Convorbiri literare; Moldova socialist; Vatra românească; Dimineața copiilor; Evenimentul zilei; Gândul; Ziua; Ziua news; Jurnalul național*
	1992-present	8,751	
Bessarabia	1941-1991	13,534	*Basarabia; Curierul; Deșteptarea; Literatură și artă; Contrafort; Jurnal; Jurnal de Chișinău; Literatură și artă; Moldova suverană; Ziarul de gardă*
	1992-present	13,471	

7.4.2 Experiments and results

We separately conducted statistical analyses on the two groups of linguistic indices calculated by *ReaderBench* were split into two primary categories:

1. *Lexical* (i.e., 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)
2. *Cohesion and Semantics* (i.e., average paragraph-document cohesion – LSA, average paragraph-document cohesion – LDA, average intra-paragraph cohesion – LDA, and average transition cohesion – LDA)

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 7.12 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.

Table 7.11 Different SNA metrics applied on both models

Index	Region	1941-1991	1992-present
Lexical indices			
Average word length in characters	Romania	3,89 (0,42)	4,13 (0,40)
	Bessarabia	3,95 (0,39)	4,38 (0,39)
Average number of unique content words per sentence	Romania	6,95 (3,09)	7,41 (2,23)
	Bessarabia	5,80 (2,40)	8,20 (1,89)
Word entropy	Romania	4,89 (0,39)	4,65 (0,20)
	Bessarabia	4,87 (0,30)	4,73 (0,46)
Average distance between lemma and word stems (only content words)	Romania	0,85 (0,16)	0,9 (0,17)
	Bessarabia	0,85 (0,16)	0,93 (0,13)
Average distance between words and corresponding stems (only content words)	Romania	1,23 (0,25)	1,35 (0,25)
	Bessarabia	1,32 (0,26)	1,44 (0,30)
Cohesion and Semantics			
Average paragraph-document cohesion (LSA)	Romania	0,63 (0,14)	0,60 (0,11)
	Bessarabia	0,59 (0,16)	0,61 (0,11)
Average paragraph-document cohesion (LDA)	Romania	0,64 (0,11)	0,63 (0,10)
	Bessarabia	0,63 (0,13)	0,61 (0,07)
Average sentence-paragraph cohesion (LSA)	Romania	0,61 (0,14)	0,68 (0,12)
	Bessarabia	0,65 (0,12)	0,72 (0,12)
Average intra-paragraph cohesion (LDA)	Romania	0,41 (0,12)	0,49 (0,12)
	Bessarabia	0,44 (0,12)	0,49 (0,07)
Average transition cohesion (LDA)	Romania	0,43 (0,10)	0,45 (0,09)
	Bessarabia	0,44 (0,14)	0,47 (0,10)

(1) ***Lexical indices***. 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.

(2) ***Cohesion and Semantics***. Because the index demonstrated the smallest effect [average paragraph-document cohesion (LSA)] we decided to remove it 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 7.15.

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 7.15a). 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 the word *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 7.15b). 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 7.15c).

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 Romanized words with Russian roots and words/sintagms derived from Russian words.

As an example, the sentence: *M-ai ubidit, tovarășe*!

is structured as follows: *M-ai* → (En.) *me*

ubidit → (Ro.) *convinge* → (En.) *convince*, derived from the Russian root, *увидит*

tovarășe → (En.) *mister*, a word assimilated with the communist regime, which has been replaced with *domnule*.

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: (eu) *Nu mă pot focusa*! → (En) I cannot focus!

Note that in Romanian, the personal pronoun can lack (e.g. eu → (En) I), being hosted by the inflected.

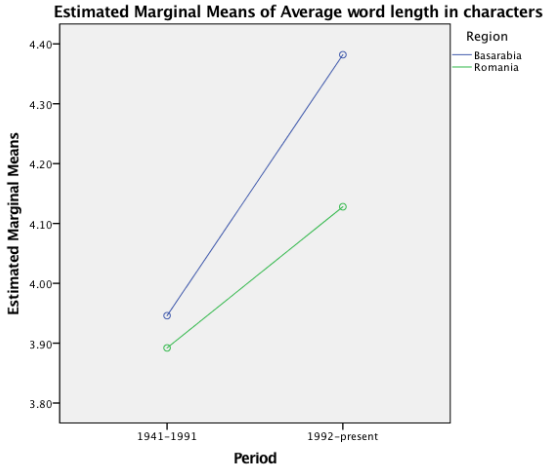
is structured as follows: *Nu* → (En.) *not*

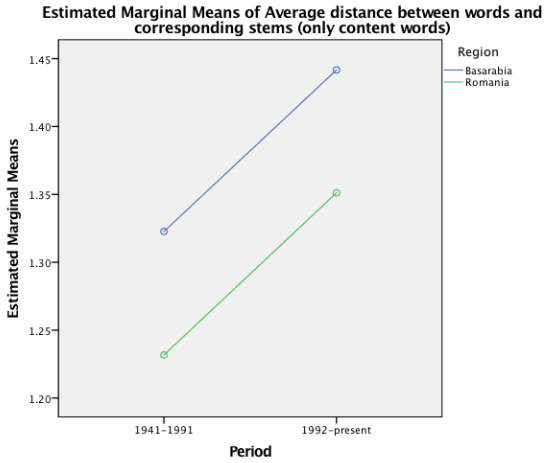
mă → (En.) *me*

pot → (En.) *can*

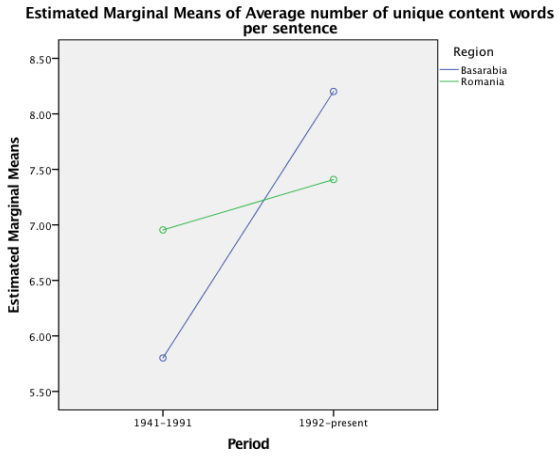
focusa → (En.) *to focus*, a Romanian new word with English root.

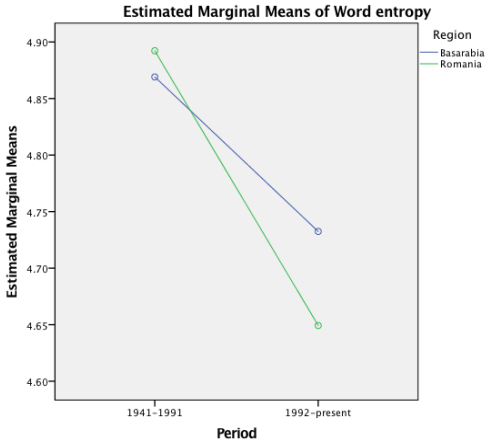
Moreover, texts tended to become more complex with longer sentences usually with 3 to 5 clauses.



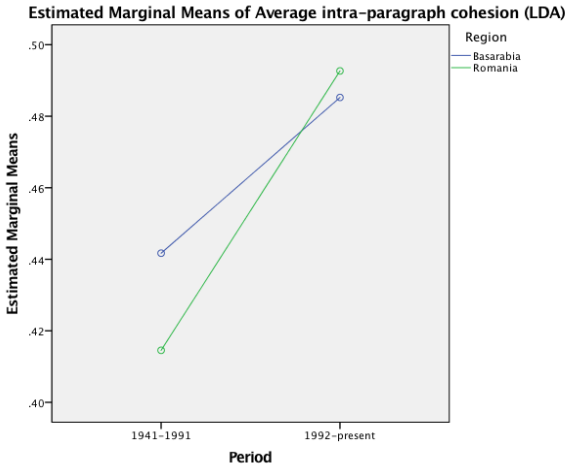


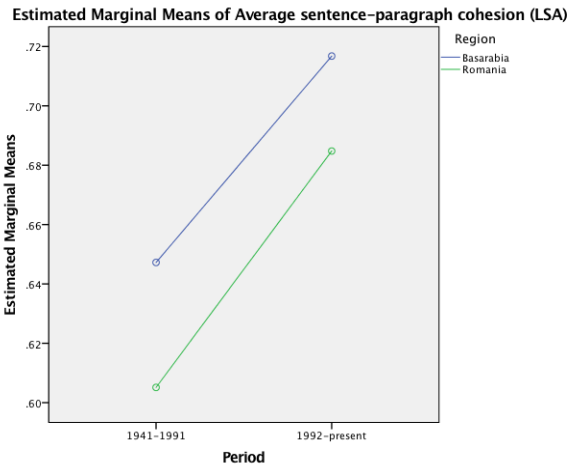
(a)





(b)





(c)

Figure 7.15 Comparative views of the Romanian language time evolution in both regions.

7.5 Conclusions

Language was not and is not static, but the feature that characterizes the language is the dynamism, whether it focuses on internal processes of word formation or loanwords. Below, we mentions a few conclusions for each module described at the current diachronic methodology.

Text Processing – We were able to successfully create a search system for unknown words (depending by DEX), acting especially on old text fields. These facts represent a premiere for the Romanian language. Two instruments of the Faculty of Computer UAIC were used, thus, proving their usefulness, and, also, improving existing findings: morphological and syntactic Tagger (WebPosRo) and Graphical Grammar Studio. This resource can be useful in other projects on the same topic. It was created an ontology of types, namely demographic time and space. The lists of words NotInDict are made with precision, being collected after a lengthy process, which was given an increased attention. Algorithms, heuristics, and techniques of implementation are independent by domain and language.

Recovering words chronoloy and old lemma using eDTLR – With the NotInDict word lists collected in the previous stage, we described a method of exploring of these old words using citations included in eDTLR. In the Romanian language study, two major objectives have been achieved using eDTLR: tracking chronology of the old words identified in the historical corpus built in this PhD thesis, using the citations from eDTLR, and recovering old lemmas considering the word-title from eDTLR, based on the lists of words NotInDict, in order to implement a diachronic POS tagger.

Language similarity – This module 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 cognate newspapers languages. In particular, we focused on the relation between Romanian (including the historical regions: Moldova, Transylvania, and Wallachia) and Bessarabian. The relation started and ended with a high level of similarity.

The methodology we described is based on statistical analysis of words distributions over epochs, reflected on the Romanian printing press, and on two classifiers, SVM and LSTM, for each epoch. The methodology presented is language independent and it 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. However, it is hardly possible to cover a

large material only by human effort necessary to improve the accuracy of the results. Out of this context, it is also important for aligned journalistic corpora with the most important Romanian language resources as WebDEX (*Romanian Academy Explanatory*), DEX on-line, and eDTLR (*Dictionary Thesaurus of the Romanian Language in electronic form).*

By using *MaxEnt* classifier, a few rules govern the differences between those four lexicons. Furthermore, we discovered correlations between rules and some common features of the words affected by them, like morphologic and etymologic traits. The results of the automatic correlation method presented are promising. They give rise to numerous further research opportunities. For example, we will create a small amount of parallel corpus to use it to first learn a kernel of REPs, and, then, to grow it using the larger non-parallel corpus. Using such a model in a contemporary Romanian POS tagger might yield interesting results, for handling unknown words. Syntax and orthography play a very important role in POS tagging, and a Gold corpora is required for training such a model.

The results of the pilot analyses, based on *ReaderBench platform,* 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 LSA and LDA. The results also are promising. This study may be developed in close interdependence with implementing of a diachronic POS-tagger. At this time, the *ReaderBench* platform responded well to the texts starting with 1941, being dependent on the Romanian language tools that we have at the moment, particularly the POS tagger for Romanian contemporary language.

8. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

When computing the variation between lexicons belongs to cognate languages, there are many aspects which can be taken into account, some of which being subjective in nature. The preliminary conclusion drawn from the previous chapters, having scrutinized various substring replacement operations at explaining the relationship between language acquisition and diachronic change in the lexical and semantics domains, confirms that language variations have a crucial role in the explanations of language evolution. Moreover, the semi-automatic recognition of language deviations represents the basis for defining the link between them, exploring the changes in the lexical and semantics rules, while comparing versions of Romanian language over time from different geographical areas. Because it is not like that: *the language is done by change and "dies" as such when it ceases to change* […]. [Coșeriu, 1997:246]

Our entire analytical approach helped us to prove that those cognate languages from Moldavia, Tranylvania, Wallachia, and Bessarabia presents many similarities between 1829-2015, identified in the current corpus built in this PhD thesis, despite the historical waves and external language influences relied on most studies of linguists.

8.1 Overview

The development of multiple diachronic models is a complex analysis in terms of the linguistic features, the multitude of considered approaches, the unitary approach that addresses to the texts from written press from four historical Romanian regions, and the focus on developing a historical corpus for Romanian language so necessary to specialists in NLP during the consideration process of language evolution in a certain time. These are just some aspects that our research reveals us.

This diachronic methodology presented here is language independent and it 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 fact, we talk about an interdisciplinary approach in which the informatics, with an accent on natural language

processing, and sciences of the language (phonetics, lexicology, grammar, semiotics) meet to complete the automatic assessments in the contrastive analysis proposed in this research.

Our diachronic methodology can be considered the most comprehensive NLP method developed so far. We refer to a complex analytic context that incorporates new ways to evaluate the language variation over time between cognates languages by using the advanced text mining techniques and models. Furthermore, these statistical models are designed considering two dimensions. On the one hand, epoch detection that highlights the significant changes in the distribution of terms. On the other hand, identifying lexical-semantic features to measure the language similarity, reflected in differences in lexical level, writing style, ranging from sentence and paragraph structure, to topic cohesion measured in terms of LDA, etc. We would like to mention the validations resulting when computing the variation between two lexicons belongs to different versions of the same language using MaxEntropy model, followed by validations focused on the relation between cognates languages. These are based on statistical analysis of words distributions over epochs belonging to a particular text written, reflected on the journalistic language and a statistical classifier, SVM, for each epoch, etc.

There were performed experiments and a verification of the profitability of each statistical model to cover a wide spectrum related to the premise of this PhD thesis, *that a deep language analysis for machine learning is best applicable in diachrony*. Moreover, all theoretical aspects argue that this diachronic methodology – we considered – may be appropriate in different contexts for the study of language.

Recently, the interest of some researchers is to predict future trends via deep learning methods with unsupervised learning on the basis of weak signals. It is very exciting and it has a lot of applications, including in diachrony. The diachronicity represents the evolution and change of certain topics in mass media over time and is linked to the detection of weak signals. We found results useful for statistical tests presented for epoch detection in [Wang & McCallum, 2006], and also in [Popescu & Strapparava, 2014].

Eventually, we believe that the initial goal to retrospectively investigate the journalistic language between related languages can be achieved using various undeniable statistical models, as in this PhD thesis, in conjunction with the reliability of the results validated in scientific events in which they were presented (LREC-2014, CICLing-2015/2016, NLPMJ held at the IJCAI-2016, ICCCI 2016, etc.).

8.2 PhD contributions

As we mentioned in the previous section, this PhD thesis presents the PhD student's work in the computational linguistics, especially in measuring language similarity, using different statistical methods in order to implement a language independent system from a diachronic perspective. The PhD student's contributions are the following:

In **Chapter 1,** the author delimits the thematic area of PhD thesis with interdisciplinary specific, being essential in the architecture of an automatic language processing system. She also describes the purpose and the major interests of addressing such topics, the temporal criterion becoming essential in capturing language phenomena in a given space.

In **Chapter 2,** the author provides a synthesis of the most important contributions in diachrony with NLP methods (historical corpora, diachronic text classification, and language similiarities using machine learning, machine translation, and metrics, epoch detection), necessary to better perceive the differences occurring in the Romanian language over time in different geographical areas.

In **Chapter 3,** the author gives a brief introduction in lexical semantics from which it appears the fact that the semantic evolution of the words is reduced to the meaning change, and its widening and restriction. Furthermore, the diachronic semantics aims to reduce the diversity of the individual meaning changes to some analogous general trends in historical phonetics.

In **Chapter 4,** the author relates a brief overview of the achievements and the lexical perspectives in language technology (in Romania), in general, and that of the language technologization in particular, with the goal to highlight the importance of the lexical resources in NLP.

In **Chapter 5,** the author briefly presented the characteristics and advantages of the most known text mining techniques and models, frequently used in various NLP applications, especially in diachrony. In our case, text mining seeks to extract useful information from a corpus (journals collections monitored over time) through the identification and exploration of relevant models.

In **Chapter 6,** the author describes the stages of the written press in Romanian language from Romania (1829-2015) and Bessarabia (1840-2015), with some details for the monitorized and representative publications in order to build a lexicon of words and phrases that will be correlated with the existing Romanian language resources as: DEX-online, RoWordNet, and e-DTLR. It is just a synthetic and abstract coverage of the historical writings about such an action, representingthe main starting point of the applied device from the next chapter.

In **Chapter 7,** the author presents the research results on the design of the diachronic methodology, which is considered the most complex method developed in Romania so far. It integrates several modules to evaluate the evolution of the Romanian language by using advanced text mining techniques and models: building historical corpus, processing text, recovering the chronological order of old Romanian words, epoch detection, language similarity computation, etc.

8.3 Directions for Future Work

Given that this research belongs to an interdisciplinary area, the directions for future research have to combine several areas, including lexicography, statistical NLP, morphosyntax and Language Semantics, etc.

From a lexicographic point of view, the first thing that needs to be done is the development of a diachronic corpus of hundreds of millions words for Romanian language, as we found for English (COHA described in [Davies, 2015], Google Books Ngram [Michel *et al.*, 2011]), for Italian (DiaCORIS presented in [Onelli *et al.*, 2006]), for Spanish (IAC detailed in [Sánchez-Marco *et al.*, 2010]), etc. reminded in Chapter 2. This requires a national mobilization of the researchers from the Institute of Linguistics, students, and professors of the Faculty of Letters from Romanian and Bessarabian universities, etc.

Secondly, a major impact in Romanian morphology would be to implement a diachronic POS Tagger, proposing a XML schema annotation of the historical corpus at the lemma level, and also using the eDTLR. As we mentioned, this type of research started a few years ago [Cristea *et al.*, 2012] by a group from Iași to whom we found a possibility to be continued in this research, but strictly in relation to this corpus, that is still in the development phase.

Thirdly, an interesting experiment could be focused on the transliteration differences from Cyrillic to Latin, both in Romania and Bessarabia until 1862, when in Romania the texts were published in both alphabets. Moreover, given that in the period 1944-1989 (excluding the war years) in Bessarabia the writing in Latin alphabet was prohibited, the process of collecting and transliterating publications of those times - with the support of the Academy of Sciences of Chisinau – should continue.

Fourthly, to provide a complex picture for language similarity between cognate languages over time, we considered that Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) best reflect the syntactically and semantically differences in writing style.

Using the two models together is a novelty in diachrony, that we have applied only on the corpus between 1941-2015, excluding the Bessarabian newspapers published in the Cyrillic alphabet. Until now, we have insufficiently old training data used by the machine learning and, implicitly, the lack of a diachronic POS-tagger for Romanian.

As a fifth research direction regarding language similarity between Romance languages (Latin, Romanian, Italian, Spanish, Catalan, French and Portuguese), we have into attention an extension of the research already known in the literature based on the syllabic similarities (Latin, Romanian, Italian, Spanish, Catalan, French and Portuguese) in different epochs, or on the orthographic similarity, presented in [Dinu & Dinu, 2005] and respectively, in [Ciobanu & Dinu, 2014].

APPENDIXES

Appendix 1 Corpus Statistics in Chronological Order

Name	Region	Year	Total NotInDict	Unique NotInDict	Total words	% (Total NotInDict /Total words)
Basarabia reinoită	Bessarabia	1817	12	10	375	3,2
Democratul Basarabiei	Bessarabia	1819	122	87	1712	6,26
Curier românesc	Wallachia	1829	11	11	207	5,31
Albina românească	Moldavia	1829	51	51	307	16,61
Albina românească	Moldavia	1832	20	20	152	13,15
Buletin. Gazeta oficială	Wallachia	1840	9	8	91	9,89
Albina românească	Moldavia	1844	41	34	410	10
Organulu Luminarei	Transylvania	1847	769	610	1994	38,56
Cuvânt Dimitrie Balica	Bessarabia	1848	241	206	695	34,67
Organulu Luminarei	Transylvania	1848	393	282	974	40,34
Romania	Wallachia	1848	21	18	265	7,92
Gazeta de Transilvania	Transylvania	1852	194	168	439	44,19
Gazeta Transilvaniei	Transylvania	1852	193	159	445	43,37
Telegrafulu Romanu	Transylvania	1853	28	28	378	7,4
Foaia pentru Minte Anima și Literatura	Transylvania	1854	667	467	1715	38,89
Federațiunea	Transylvania	1856	119	102	255	46,66
Foaia pentru Minte Anima și Literatura	Transylvania	1860	552	419	1622	34,03
Telegraful roman	Wallachia	1865	2109	1283	4475	47,12

Foaia pentru Minte Anima și Literatura	Transylvania	1865	488	323	1227	39,77
Transilvania	Transylvania	1866	85	70	516	16,47
Convorbiri literare	Moldavia	1868	1417	981	5601	25,29
Federațiunea	Transylvania	1868	2035	1331	4668	43,59
Transilvania	Transylvania	1869	779	411	2031	38,35
Albina romanească	Moldavia	1869	234	187	546	42,85
Transilvania	Transylvania	1869	185	138	472	39,19
Federațiunea	Transylvania	1869	64	44	153	41,83
Gura Satului	Transylvania	1869	32	29	76	42,1
Convorbiri literare	Moldavia	1870	65	53	327	19,87
Albina	Transylvania	1870	589	452	1405	41,92
Federațiunea	Transylvania	1870	114	103	337	33,82
Gura Satului	Transylvania	1870	125	97	283	44,16
Albina	Transylvania	1870	118	108	1405	8,39
Telegraful Romanu	Transylvania	1873	102	89	496	20,56
Curierul românesc	Wallachia	1875	84	69	374	22,45
Telegrafulu romanu	Transylvania	1875	3307	1911	8152	40,56
Curierul	Bessarabia	1876	170	126	730	23,28
Transilvania	Transylvania	1876	310	211	782	39,64
Pressa	Wallachia	1877	493	325	5436	9,06
România liberă	Wallachia	1877	329	255	3876	8,48
Romanulu	Wallachia	1877	340	237	4789	7,09
Timpul	Wallachia	1877	20	19	2027	0,98
Telegraful roman	Transylvania	1877	3354	1814	8009	41,87
Pressa	Wallachia	1878	747	414	7562	9,87
Romania liberă	Wallachia	1878	388	263	3300	11,75
Romanulu	Wallachia	1878	433	247	6493	6,66
Timpul	Wallachia	1878	294	232	7273	4,04
Literatorul	Wallachia	1880	412	307	1932	21,32

Pressa	Wallachia	1880	674	420	6520	10,33
România liberă	Wallachia	1880	604	409	5549	10,88
Romanulu	Wallachia	1880	345	216	5605	6,15
Timpul	Wallachia	1880	898	526	19304	4,65
Curierul. Foaia intereselor generale	Moldavia	1880	84	72	3795	2,13
Gura Satului	Transylvania	1880	398	287	984	40,44
Familia	Transylvania	1881	368	267	2053	17,92
Gura Satului	Transylvania	1881	408	330	999	40,84
Timpul	Wallachia	1883	137	112	918	14,92
Telegraful roman	Transylvania	1883	81	59	515	15,72
Albina	Wallachia	1883	33	31	146	22,6
Albina	Wallachia	1884	41	38	446	9,19
Constituționalul	Moldavia	1885	136	118	539	25,23
Transilvania	Transylvania	1887	144	117	429	33,56
Telegraful roman	Transylvania	1888	807	499	2836	25,83
Telegraful roman	Transylvania	1889	232	189	1020	23,17
Candela	Bessarabia	1890	185	122	582	31,78
Familia	Transylvania	1890	100	52	433	23,09
Telegraful roman	Transylvania	1890	85	74	383	22,19
Aradu	Transylvania	1890	2584	1594	6457	40,01
Convorbiri literare	Moldavia	1893	4348	2107	38267	11,36
Familia	Transylvania	1894	35	28	702	4,98
Telegraful roman	Transylvania	1895	18	15	84	21,42
Desteptarea	Bessarabia	1896	105	67	878	11,95
Convorbiri literare	Moldavia	1897	126	92	2579	4,88
Basarabia reinoită	Bessarabia	1907	74	63	1137	6,5
Desteptarea. Foaie pentru popor	Wallachia	1911	16	13	595	2,68
Cuvânt moldovenesc	Bessarabia	1913	181	118	1073	16,86
Curierul	Bessarabia	1918	39	27	816	4,77

Ardealul	Bessarabia	1918	26	20	554	4,69
Basarabia	Bessarabia	1918	16	12	250	6,4
Romania nouă	Bessarabia	1918	12	9	644	1,86
Sfatul țării	Bessarabia	1918	242	171	4125	5,23
Basarabia	Bessarabia	1919	14	12	293	4,77
Democratul Basarabiei	Bessarabia	1919	218	142	3531	6,17
Glasul Basarabiei	Bessarabia	1919	14	14	149	9,39
Luminătorul	Bessarabia	1919	37	20	658	5,62
Adeverul	Wallachia	1919	97	73	1596	6,07
Curierul artelor	Wallachia	1919	78	62	1675	4,65
Dimineata	Wallachia	1919	25	17	406	6,15
Universul	Wallachia	1919	56	42	1010	5,54
Patria	Transylvania	1919	234	160	4642	5,04
Moldova Socialistă	Moldavia	1919	83	65	2408	3,44
Scânteia	Moldavia	1919	123	104	2224	5,53
Telegraful roman	Transylvania	1919	7	6	115	6,08
Sfatul țării	Bessarabia	1920	131	112	2151	6,26
Universul	Wallachia	1920	11	10	114	9,64
Dreptatea	Bessarabia	1922	3	3	102	2,94
Luminătorul	Bessarabia	1923	2	1	139	1,43
Luminătorul	Bessarabia	1924	31	29	556	5,57
Luminătorul	Bessarabia	1925	14	11	209	6,69
Adeverul	Wallachia	1925	1	1	60	1,66
Universul	Wallachia	1925	15	14	308	4,87
Dreptatea	Bessarabia	1926	62	51	1390	4,46
Viitorul	Wallachia	1926	19	15	271	7,01
Chemarea tinerimei romane	Transylvania	1926	55	50	1488	3,69
Dreptatea	Transylvania	1926	14	14	468	2,99
Solidaritatea	Bessarabia	1927	8	8	384	2,08

Luminătorul	Bessarabia	1928	35	28	886	3,95
Curentul	Wallachia	1928	127	84	2525	5,02
Dimineata	Wallachia	1928	22	15	355	6,19
Basarabia	Bessarabia	1929	45	40	775	5,8
Glasul Basarabiei	Bessarabia	1929	15	15	452	3,31
Luminatorul	Bessarabia	1929	22	18	527	4,17
Dreptatea	Transylvania	1929	56	30	1908	2,93
Curentul	Wallachia	1929	30	13	386	7,77
Universul	Wallachia	1929	21	19	252	8,33
Luminatorul	Bessarabia	1930	16	12	458	3,49
Universul literar	Wallachia	1930	15	12	138	10,86
Aradul	Transylvania	1930	18	17	268	6,71
Solidaritatea	Bessarabia	1930	98	64	2583	3,79
Basarabia Chișinăului	Bessarabia	1931	25	21	204	12,25
Dimineata	Wallachia	1931	2	2	132	1,51
Aradul	Transylvania	1931	29	21	304	9,53
Basarabia	Bessarabia	1932	77	58	1710	4,5
Curierul	Bessarabia	1932	53	44	1676	3,16
Sfatul țării	Bessarabia	1932	20	18	292	6,84
Curentul	Wallachia	1932	17	15	681	2,49
Solidaritatea	Bessarabia	1932	7	7	200	3,5
Luminatorul	Bessarabia	1933	18	16	525	3,42
Adevarul	Wallachia	1933	10	6	100	10
Curentul	Wallachia	1933	2	2	77	2,59
Universul	Wallachia	1933	6	6	199	3,01
Curierul crestin	Transylvania	1933	26	19	411	6,32
Libertatea. Organ politic muncitoresc	Moldavia	1933	28	17	439	6,37
Luminatorul	Bessarabia	1934	14	13	672	2,08
Universul	Wallachia	1934	7	5	149	4,69

Curierul crestin	Transylvania	1934	25	21	442	5,65
Contabilul	Moldavia	1934	9	7	261	3,44
Solidaritatea	Bessarabia	1934	13	13	279	4,65
Luminatorul	Bessarabia	1935	13	11	396	3,28
Adeverul	Wallachia	1935	80	50	1680	4,76
Curentul	Wallachia	1935	5	4	88	5,68
Familia	Transylvania	1935	95	85	533	17,82
Luminatorul	Bessarabia	1936	80	65	2460	3,25
Universul	Wallachia	1936	21	18	490	4,28
Adevarul literar si artistic	Wallachia	1937	5	3	163	3,06
Curentul	Wallachia	1937	19	14	481	3,95
Basarabia	Bessarabia	1938	9	8	97	9,27
Curentul	Wallachia	1938	2	2	85	2,35
Universul	Wallachia	1938	84	61	1803	4,65
Luminatorul	Bessarabia	1939	14	10	703	1,99
Curierul	Wallachia	1939	39	29	864	4,51
Desteptarea	Moldavia	1940	6	5	87	6,89
Curentul	Wallachia	1940	10	8	706	1,41
Solidaritatea	Bessarabia	1940	19	18	791	2,4
Basarabia	Bessarabia	1941	38	32	1335	2,84
Luminatorul	Bessarabia	1941	13	12	568	2,28
Buletinul Arhiepiscopiei Chișinăului	Bessarabia	1942	9	9	326	2,76
Viata economică din Bălți	Bessarabia	1943	8	7	275	2,9
Buletinul Arhiepiscopiei Chișinaului	Bessarabia	1944	1	1	73	1,36
Discurs Onufrie Andrus RSSM	Bessarabia	1945	3	2	46	6,52
Ethos	Bessarabia	1945	5	5	256	1,95
Scânteia	Wallachia	1947	1	1	153	0,65
Vatra românească	Transylvania	1949	44089	16320	860871	5,08
Vatra românească	Transylvania	1950	86450	31616	1359366	6,29

Sania	Bessarabia	1955	111	34	2840	3,9
Programe pentru școala medie din Chișinău	Bessarabia	1962	3	3	324	0,92
Moldova socialistă	Bessarabia	1967	240	182	4081	5,97
Rev. Univ. Chișinău	Bessarabia	1968	22	16	1024	2,14
Cuvînt Ion Druță	Bessarabia	1970	4	1	561	0,71
Povara bunătății noastre	Bessarabia	1970	193	71	13170	1,46
Echinox	Transylvania	1973	27	24	970	2,78
Echinox	Transylvania	1974	28	23	2098	1,33
Viata satului	Bessarabia	1975	50	37	367	13,62
Viata satului	Bessarabia	1983	29	14	879	3,29
Echinox	Transylvania	1986	1	1	66	1,51
Femeia Moldovei	Bessarabia	1986	155	106	5371	2,75
Deșteptarea	Bessarabia	1989	6	6	457	1,31
Literatura și artă	Bessarabia	1989	17	13	610	2,78
Romania literară	Wallachia	1989	41	32	2028	2,02
Scânteia	Wallachia	1989	8	8	1796	0,44
Convorbiri literare	Moldavia	1989	26	19	1800	1,44
Curierul	Bessarabia	1990	48	30	2282	2,1
Deșteptarea	Bessarabia	1990	55	30	1532	3,59
Literatura și artă	Bessarabia	1990	49	44	1324	3,7
Deșteptarea	Moldavia	1990	27	24	970	2,78
România literară	Wallachia	1990	111	85	3289	3,37
Convorbiri literare	Moldavia	1990	9	3	401	2,24
Moldova Socialistă	Moldavia	1990	20	19	904	2,21
Curierul	Bessarabia	1991	108	68	2790	3,87
Deșteptarea	Bessarabia	1991	4	2	281	1,42
Literatura și artă	Bessarabia	1991	42	32	2122	1,97
România literară	Wallachia	1991	25	24	793	3,15
Convorbiri literare	Moldavia	1991	137	117	4359	3,14

Dimineata copiilor	Wallachia	1993	40	15	584	6,84
Literatură și artă	Bessarabia	1998	8	6	219	3,65
Buna ziua Iasi	Moldavia	1999	69	60	437	15.78
Adevarul de Cluj	Transylvania	2003	45	34	3181	1.41
Faclia de Cluj	Transylvania	2004	12	10	1070	1.12
Faclia de Cluj	Transylvania	2008	14	4	581	2.4
Jurnal	Bessarabia	2009	64	33	886	7,22
Ziarul de Vrancea	Moldavia	2009	59	48	379	15.56
Gazeta de Transilvania	Transylvania	2009	2	2	232	0.86
Jurnal	Bessarabia	2010	36	16	963	3,73
Gazeta de Transilvania	Transylvania	2010	7	6	342	2.04
Monitorul de Vaslui	Moldavia	2011	14	12	228	6.14
Evenimentul regional al Moldovei	Moldavia	2012	12	10	549	2.18
Monitorul de Cluj	Transylvania	2012	19	15	560	3.39
Contrafort	Bessarabia	2013	7	4	217	3,22
Jurnal de Chișinău	Bessarabia	2013	40	9	1432	2,79
Buna ziua Iasi	Moldavia	2013	28	20	166	16.86
Contrafort	Bessarabia	2014	25	18	1025	2,43
Moldova suverană	Bessarabia	2014	521	184	12097	4,3
Ziarul de gardă	Bessarabia	2014	630	229	17163	3,67
Evenimentul zilei	Wallachia	2014	275	133	7383	3,72
Gândul	Wallachia	2014	675	208	15988	4,22
Ziua	Wallachia	2014	186	127	4280	4,34
Ziua news	Wallachia	2014	14	3	525	2,66
Ziua veche	Wallachia	2014	374	200	6358	5,88
Impartial	Moldavia	2014	2	2	113	1.76
Literatură și artă	Bessarabia	2015	51	28	926	5,5
Literatura si arta	Bessarabia	2015	51	28	926	5.5
Ziarul de Vrancea	Moldavia	2015	22	18	137	16.05

Bihoreanul	Transylvania	2015	4	4	432	0.92

Appendix 2 Chronological order of the old words according with eDTLR (a sample)

Our corpus				eDTRL			
NotInDict	**Year**	**Region**	**Article**	**Year**	**Quotations**	**Sigla**	**Title**
aceastea	1918	Bessarabia	Sfatul țării	1581	Laban răspunse și zise lui Jacov: **aceastea** feate, featele mele.	PALIA	Palia d'Orăştie. (1581–1582). I: Préface et livre de la Genèse, publiés avec le texte hongrois de Heltai et une introduction par Mario Roques. Les premières traductions roumaines de l'Ancien Testament. Paris, Champion, 1925.
acelu	1847	Transylvania	Organulu Luminarei	1574	Că **acelu** lemn plecat iaste și scundicel și iaste în chipul smochinului.	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
acestă	1880	Wallachia	Romanulu	1575	Cu destoinicie păţim aceasta, c-am greşit înaintea fratelui nostru şi vădzînd spaima sufletului său, cîndu se milcuiia noao, nu vrum noi să-l ascultăm, derept aceaia vine **acestă** dosadă pre noi.	PO	P. Poenaru, Elemente de geometrie, după Legendre. Bucureşti, Tipografia lui Eliad, 1837.

aducu	1868	Moldavia	Convorbiri literare	1521	Chiemă-me de me rugă se **aducu** acestu giure la tinre.	ib.	
Be	1950	Transylvania	Vatra romaneasca	1645	**Be** și să disfăta, giucînd și șuguind cu tovarășii.	HERODOT	Herodot, Istorii. [Traducere de Nicolae Spătarul]. Ediție îngrijită de Liviu Onu şi Lucia Şapcaliu. Prefaţă, studiu filologic, note, glosar de Liviu Onu. Indice de Lucia Şapcaliu. Bucureşti, Editura Minerva, 1984 (Ediţii critice).
blăstămuri	1934	Bessarabia	Luminătorul	1693	Cu mari **blăstămuri** și sudalme pe Pardos și pe Corb încărcînd.	CANTEMIR, I. I.	Dimitrie Cantemir, Istoria ieroglifică. [Vol.] I–II. Bucureşti, Editura pentru Literatură, 1965.
Baiazid	1880	Moldavia	Curierul. Foaia intereselor generale	1783	**Baiazid**… luă vestire că Sigismond…au trecut Dunărea.	VĂCĂRESCUL, IST.	Ienăchiță Văcărescu, Istoria othmanicească. Ediție critică, studiu introductiv, note şi glosar de Gabriel Ştrempel. [Bucureşti], Editura Biblioteca Bucureştilor, 2001.
Băcăului	1878	Wallachia	Timpul	1798	Tătarii ... au dat în urma oștilor celor protivnice și, agiungându-se în Moldova la pădurea **Băcăului**, dânsele n-au stat la bătaie.	ȘINCAI, HR.	

cadi	1868	Moldavia	Convorbiri literare	1784	Copaciul mari cându **cadi** mari urlet faci.	ALEXANDRIA	Istoria lui Alexandru cel Mare. A cincea ediție. Bucureşti, tipărită la Iosef Copainig, 1850.
caile	1883	Wallachia	Timpul	1600	Într-aceaia vreme închise vrăjmașul toate **caile** să nu meargă oaspe la casa lui Avram	CUV. D. BĂTR.	B. Petriceicu-Hasdeu, Cuvente den bătrâni. Tomul I. Limba română vorbită între 1550–1600. Studiu paleografico-linguistic de B. Petriceicu Hasdeu...; Tomul II. Cărţile poporane ale românilor în secolul XVI … ; Tomul III. Istoria limbei române. Partea I. Principie de linguistică. Ediție îngrijită, studiu introductiv şi note de G. Mihăilă. Bucureşti, Editura Didactică şi Pedagogică, [vol.] I–III, 1983-1987.
cântu	1950	Transylvania	Vatra romanească	1550	Laudă şi strigare **cântu** şi cânta-voiu Domnului.	PSALT.	Psaltirea Hurmuzaki. [cca. 1500–1510] [Manuscris de la începutul secolului al XVI-lea; Biblioteca Academiei Române, cota 3077].
câ	1913	Bessarabia	Cuvânt moldovenesc	1575	Cetiți şi nu judccareți necetind mainte, **câ** veți cu adevăt a afla mare vistiiariu sufletesc.	PO	P. Poenaru, Elemente de geometrie, după Legendre. Bucureşti, Tipografia lui Eliad, 1837.
da'	1990	Moldavia	Convorbiri literare	1646	Ia vezi, că jupînu tace, tace... **da'** cînd o-ncepe o dată, apoi parcă vă văz iar cu vînătăile pe subt ochi.	PR. DRAM.	Primii noştri dramaturgi. Ediţie îngrijită şi glosar de Al. Niculescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, [1956]. (Clasicii români).
darulu	1890	Transylvania	Aradu	1700	Pentru **darulu** înțelepției spun la istoriile Rîmului că au încălecat într-o zi împăratul şi află într-o luncăminica67059pre un filosof singur	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
decâtu	1875	Wallachia	Curierul romanesc	1644	Să nu ne arătăm istorici de lucruri turceşti **decâtu** de ale noastre.		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.

dece	1809	Bessarabia	Democratul Basarabiei	1649	Îmblă neguțătorind pre la casele muerilor, **dece** le dă tot mai eftin de cumu-i prețul.	MARDARIE, L.	Mardarie Cozianul, Lexicon slavo-românesc şi tâlcuirea numelor din 1649. Publicate cu un studiu, note şi indicele cuvintelor româneşti de Grigorie Creţu. Cu mai multe facsimile. Ediţiunea Academiei Române. Bucureşti, Carol Göbl, 1900.
Eco	2014	Bessarabia	Ziarul de garda	1693	Dragoș Vodă au spus acelui **Eco** cum că iaste vechi	N. COSTIN, L.	Nicolae Costin, Opere. [Vol.] I: Letopiseţul Ţării Moldovei de la zidirea lumii până la 1601 şi de la 1709 la 1711. Ediţie cu un studiu introductiv, note, comentarii, indice şi glosar de Const. S. Stoide şi I. Lăzărescu. Cu prefaţă de G. Ivănescu, Iaşi, Editura Junimea, 1976.
eftin	1950	Transylvania	Vatra romaneasca	1649	Îmblă neguțătorind pre la casele muerilor, dece le dă tot mai **eftin** de cumu-i prețul.	MARDARIE, L.	Mardarie Cozianul, Lexicon slavo-românesc şi tâlcuirea numelor din 1649. Publicate cu un studiu, note şi indicele cuvintelor româneşti de Grigorie Creţu. Cu mai multe facsimile. Ediţiunea Academiei Române. Bucureşti, Carol Göbl, 1900.
Eghipet	1829	Moldavia	Albina romaneasca	1550	Puse în **Eghipet** semnele (znameniele C	PSALT.	Psaltirea Hurmuzaki. [cca. 1500–1510] [Manuscris de la începutul secolului al XVI-lea; Biblioteca Academiei Române, cota 3077].
elu	1865	Wallachia	Telegraful roman	1574	Să se încredințeze că era grăită de **elu** … cu această znamenie.	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
făcăndu	1950	Transylvania	Vatra romaneasca	1521	Plodul dereptației întru împăcare seamără-se, **făcăndu** pace.	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
fără	1913	Bessarabia	Cuvant moldovenesc	1646	Muiarea poate să se desparță de bărbatul său ... cându o va bate des și **fără** de vină.	PRAV.	Carte românească de învăţătură de la pravilele împărăteşti şi de la alte giudeaţe, cu zisa şi cu toată cheltuiala lui Vasilie Voivodul. Iaşi, l646.
făgădueli	1880	Wallachia	Romania libera	1783	Pe unii dobândindu-i cu **făgădueli**, pe alții silindu-i cu înfricoșări.	VĂCĂRESCUL, IST.	Ienăchiţă Văcărescu, Istoria othmanicească. Ediţie critică, studiu introductiv, note şi glosar de Gabriel Ştrempel. [Bucureşti], Editura Biblioteca Bucureştilor, 2001.
fun	1893	Moldavia	Convorbiri literare	1646	Nu știi vro odaie să-ți spun ceva d-ale lumii și să ne jucăm **fun** joc?	PR. DRAM.	Primii noştri dramaturgi. Ediţie îngrijită şi glosar de Al. Niculescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, [1956]. (Clasicii români).

gălceava	1880	Wallachia	Romania libera	1676	Atuncea s-au fost sculat vândzătorii de acolo cu **gălceava** spre Ghinea medelnicearіul	BUL. COM. IST.	Buletinul Comisiei istorice a României, publicat de Ioan Bogdan (vol. I–II), apoi de N. Iorga. Bucureşti, Atelierele Grafice Socec. Vol. I (1915) ş. u..
giudeci	1950	Transylvania	Vatra romaneasca	1521	Nu gicereți Dzeului nedereptate. PSALT, 150. Și tu ședzi de **giudeci** după leage și prespre leage treci, că gici se me bată?	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
gre	1832	Moldavia	Albina romaneasca	1784	Vineri, de să va întîmpla naștere lui Hristos, iarna va fi **gre** și năprasnică		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
greșala	1932	Bessarabia	Curierul	1640	Dumnedzău de iartă **greșala** ta, îț[i] iartă o mie de zloți.	VARLAAM, C.	Varlaam, mitropolitul Moldovei, Cazania (1643). [ed. de J. Byck]. Bucureşti, Editura Fundaţiilor, 1943. (Scriitori români vechi).
hainile	1974	Transylvania	Echinox	1688	Și fu deaca citi împăratul lui Israil cartea, rupse **hainile** lui și zise: Au D[u]mnezeu sînt eu, a învia au a omorî, de trimite acesta cătră mine om ca să-l curățescu de stricăciunea lui? Numai ce cunoașteț dară și vedeț că vinuiaște acesta pre mine.	BIBLIA	Monumenta linguae Dacoromanorum. Biblia. 1688. Pars I: Genesis, Iaşi, 1988, Pars II: Exodus, Iaşi, 1991, Pars III: Leviticus, Iaşi, 1993; Pars IV: Numeri, Iaşi, 1994; Pars V: Deuteronomium, Iaşi, 1997 (Universitatea „Al. I. Cuza”-Iaşi, Albert-Ludwigs-Universität–Freiburg).
hărtii	1880	Wallachia	Romania libera	1646	Uneori scriu **hărtii** cu sudalme și cu ocări asupra cuiva și le aruncă pre ulițe.	PRAV.	Carte românească de învăţătură de la pravilele împărăteşti şi de la alte giudeaţe, cu zisa şi cu toată cheltuiala lui Vasilie Voivodul. Iaşi, l646.
ierbile	1950	Transylvania	Vatra romaneasca	1574	Cum zboară acealea [albinele] pre câmpure și den toate **ierbile**.	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).

Isus	1930 1919	Wallachia Moldavia	Universul literar Scânteia	1640	**Isus**... întrebă-l pre îns de dzise: „Ce veri să-ți dau?“ Iară el dzise: „Doamne, să văd̆z!“	VARLAAM, C.	Varlaam, mitropolitul Moldovei, Cazania (1643). [ed. de J. Byck]. Bucureşti, Editura Fundaţiilor, 1943. (Scriitori români vechi).
îmflă	1970	Bessarabia	Povara bunătății noastre	1675	Ca o corabie cu multe veatrele când o **îmflă** zefirul, de către cârmă.	DOSOFTEI, V. S.	Mitropolitul Dosoftei, Viaţa şi petreacerea svinţilor. Iaşi, 1682.
încungiure	1880	Wallachia	Romania libera	1715	Să meargă cu oștile să **încungiure** pe de ceea parte, ca nu cumva să vie niscai oști de undeva făr’ de veste	ARHIVA	Arhiva românească. Subt redacţia lui M. Kogălniceanu. Iaşi, tom. I–II, 1841–1845.
împlă	1949	Transylvania	Vatra romaneasca	1574	Zicerea lu David să **împlă**, unde zice: „cu mâinile meale Dumnezeu istipii şi nu mă înşelai“.	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
începătoriu	1893	Moldavia	Convorbiri literare	1675	Mîntuitoriul... au vrednicit să-l aibă [pe Avram]... şi rudeniei **începătoriu** celor dintru dînsul.	DOSOFTEI, V. S.	Mitropolitul Dosoftei, Viaţa şi petreacerea svinţilor. Iaşi, 1682.
jidoviloru	1890	Transylvania	Aradu	1521	E **jidoviloru** și eleniloru sile nu proaste făcea Dumnedzeu cu mârule lu Pavelu.	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.

jumatate	2014	Wallachia	Ziua veche	1719	În capul viii din deal şi în capul viii din vale pusă viţă tînără ca **jumatate** pogon	URICARIUL	Theodor Codrescu, Uricariul cuprinzetoriu de hrisoave, anaforale şi alte acte din suta a XV– XIX, atingătoare de Moldova. Sub redacţiunea dsale... Vol. I–XXV. Iaşi, 1852–1895; pentru vol. I, II, III, v. şi ediţiunea a 2-a, Iaşi, 1871–1893.
lacrâmi	1949	Transylvania	Vatra romaneasca	1640	Aceştia... nu lasă să să rădăcineadze cuvântul luighaja23998 Dumnedzău într-înşii. Şi sămt seci, de n-au umegiune de umilinţă, sau de **lacrâmi**, sau de smerenie, ca o piatră.	VARLAAM, C.	Varlaam, mitropolitul Moldovei, Cazania (1643). [ed. de J. Byck]. Bucureşti, Editura Fundaţiilor, 1943. (Scriitori români vechi).
lemnu	1844	Moldavia	Albina romaneasca	1693	Au dus multă somă şi mare **lemnu** şi greu … cu mare grabă şi cu mulţi zapcii.	NECULCE, L.	Ion Neculce, Letopiseţul Ţării Moldovei şi O samă de cuvinte. Text stabilit, glosar, indice şi studiu introductiv de Iorgu Iordan. Ediţia a II-a, revăzută. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1959.
locu	1878	Wallachia	Timpul	1688	Au adunat prepelițele … ș-au zvântat **locu**.	BIBLIA	Monumenta linguae Dacoromanorum. Biblia. 1688. Pars I: Genesis, Iaşi, 1988, Pars II: Exodus, Iaşi, 1991, Pars III: Leviticus, Iaşi, 1993; Pars IV: Numeri, Iaşi, 1994; Pars V: Deuteronomium, Iaşi, 1997 (Universitatea „Al. I. Cuza”-Iaşi, Albert-Ludwigs-Universität–Freiburg).
lipsescu	1876	Bessarabia	Curierul	1694	Cât de ascunsu iaste un ghiuluș, ce nu-i **lipsescu** vrăjmașii de casă, carii vădesc lucratele lui	FN	Emil Vârtosu, Foletul novel. Calendarul lui Constantin-Vodă Brâncoveanu. 1693–1704. Bucureşti, 1942.
macar	1913	Bessarabia	Cuvant moldovenesc	1691	Ce zic eu pentru împăratul, **macar** de-am putea pentru sufletele noastre să ne nevoim.	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).

mainainte	1847	Transylvania	Organulu Luminarei	1688	Știia **mainainte** ale lor și cele viitoare.	BIBLIA	Monumenta linguae Dacoromanorum. Biblia. 1688. Pars I: Genesis, Iaşi, 1988, Pars II: Exodus, Iaşi, 1991, Pars III: Leviticus, Iaşi, 1993; Pars IV: Numeri, Iaşi, 1994; Pars V: Deuteronomium, Iaşi, 1997 (Universitatea „Al. I. Cuza”-Iaşi, Albert-Ludwigs-Universität–Freiburg).
maiu	1829	Wallachia	Curier romanesc	1671	Soli au trimis Ștefan Vodă cu cărți iscălite și cu pecetea sa, în care se adeveria cum la anul viitoriu, în luna lui **maiu**, va merge la craiul. URECHE, LET	M. COSTIN, O.	Miron Costin, Opere. Ediţie critică cu un studiu introductiv, note, comentarii, variante, indice şi glosar de P. P. Panaitescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1958.
mene	1893	Moldavia	Convorbiri literare	1640	Hiecarele ce va vre să vie după **mene**... să să lepede de sine și să-ș ia crucea sa și să vie după mene.	VARLAAM, C.	Varlaam, mitropolitul Moldovei, Cazania (1643). [ed. de J. Byck]. Bucureşti, Editura Fundaţiilor, 1943. (Scriitori români vechi).
n’are	1935	Bessarabia	Luminatorul	1646	De **n’are** avea bărbatul altă nici o vină, ce numai aceasta vrăjmășia, cându-i jestoc și mânios.	PRAV. MOLD.	Carte românească de învăţătură de la pravilele împărăteşti şi de la alte giudeaţe, cu zisa şi cu toată cheltuiala lui Vasilie Voivodul. Iaşi, l646.
naravurile	1865	Transylvania	Telegraful roman	1582	Să părăsească **naravurile** trupului, cealea realele, trufiile și mîndriile.	COD. TOD.	Dr. Nicolae Drăganu, Două manuscripte vechi. Codicele Todorescu şi Codicele Marţian. [secolul XVII-lea] Studiu şi transcriere de... Ediţiunea Academiei Române. Bucureşti, Librăriile Socec et. Comp. şi C. Sfetea; Leipzig, Otto Harrassowitz; Wiena, Gerold et Comp., 1914.
nepuind	1878	Wallachia	Timpul	1693	În mațe de atâta vreame cevași măcară **nepuind**..., mare iușurime trupului... să-m fie vinit.	CANTEMIR, IST.	Operele principelui Dimitrie Cantemir. Tipărite de Societatea Academică Română. Tom.VI: Istoria ieroglifică. (Operă originală inedită scrisă în limba românească la 1704). Bucureşti, 1883. [Editată de A. I. Odobescu].
năs	1893	Moldavia	Convorbiri literare	1582	Te poți rădzîma di **năs**.	COD. TOD.	Dr. Nicolae Drăganu, Două manuscripte vechi. Codicele Todorescu şi Codicele Marţian. [secolul XVII-lea] Studiu şi transcriere de... Ediţiunea Academiei Române. Bucureşti, Librăriile Socec et. Comp. şi C. Sfetea; Leipzig, Otto Harrassowitz; Wiena, Gerold et Comp., 1914.

obrazu	1880	Transylvania	Gura Satului	1521	H[risto]s chinui după noi, voao lăsă **obrazu**, cum se slediți urmelor lui.	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
oiei	1865	Wallachia	Telegraful roman	1671	Cum apucă domniia îndată să dezbrăcă lupul de pielea **oiei** și-ș astupă urechile ca aspida și ca vasiliscul.	GAVRIL, NIF.	Gavriil Protul, Viaţa şi traiul sfinţiei sale părintelui nostru Nifon, patriarhul Ţarigradului... în Literatura română veche (1402–1647). Introducere, ediţie îngrijită şi note de G. Mihăilă şi Dan Zamfirescu. Vol. I, Editura Tineretului, [1969], 66–99.
ome	1893	Moldavia	Convorbiri literare	1574	Tu însuți nu asculți rugăciunea ta, **ome**, iară Dumnezeu tu veri să o auză?	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
obiceiu	1931	Bessarabia	Basarabia Chisinaului	1799	Sosi ziua cea orânduită și se făcu nunta, mearseră în biserica Sfântului Vonifatie, și-i cununară după **obiceiu**, cu alăute, cu tămpene și cu organe și cu alte feliuri de zicături	REV. IST.	Revista pentru istorie, arheologie şi filologie, sub direcţiunea lui Gregoriu G. Tocilescu. Bucureşti. Anul I (1893) ş.u.
parinte	1875	Transylvania	Telegrafulu romanu	1649	Pătru Vodă cel de apoi [Șchiopul] cînd au fostu episcop la Roman Agafton, de postrig de Rîșga [Rîșca] și au datu cel **parinte**	IORGA, S. D.	N. Iorga, Studii şi documente cu privire la istoria românilor. Vol. I–XXXI. Bucureşti, Editura „Minerva”, 1901–1916.
pămăntului	1877	Wallachia	Romania libera	1582	Și spuș voao ce-i pre suptu cer și spre fața **pămăntului**.	COD. TOD.	Dr. Nicolae Drăganu, Două manuscripte vechi. Codicele Todorescu şi Codicele Marţian. [secolul XVII-lea] Studiu şi transcriere de... Ediţiunea Academiei Române. Bucureşti, Librăriile Socec et. Comp. şi C. Sfetea; Leipzig, Otto Harrassowitz; Wiena, Gerold et Comp., 1914.
peire	1918	Bessarabia	Ardealul	1675	Pentru pizmași îmi vei da știre Cînd le-a veni vremea de **peire**.		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.

perdu	1868	Moldavia	Convorbiri literare	1644	Iarăși, **perdu** Iliiașu Vodă războiul, cum să zice și la Scrisoare: „Unde nu va Dumnezeu, nu poate omul“.	URECHE, L.	Grigore Ureche, Letopiseţul Ţării Moldovei. Ediţie îngrijită, studiu introductiv, indice şi glosar de P. P. Panaitescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1955.
rândul	1880	Wallachia	Romania libera	1617	Iar de **rândul** celor oşti de maiinte, i-au bătut moldoveanii	ROSETTI, B.	Alexandre Rosetti, Lettres roumaines de la fin du XVI-e et du début du XVII-e siècle tirées des archives de Bistritza (Transylvanie) publiées par... Bucureşti, Socec et Co., 1926. (Institutul de Filologie şi Folclor).
răpejunea	1870	Moldavia	Convorbiri literare	1671	Pre alții i-au dus **răpejunea** apei.	M. COSTIN	Miron Costin, Opere. Ediţie critică cu un studiu introductiv, note, comentarii, variante, indice şi glosar de P. P. Panaitescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1958.
răveneala	1950	Transylvania	Vatra romaneasca	1574	Vrea să ungă trupul lui... că mirul are oarece tărie și usucă nini69380 și **răveneala** trupului suge-o.	CORESI, EV.	Diaconul Coresi, Carte cu învățătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
rimlean	1913	Bessarabia	Cuvant moldovenesc	1648	Pre acest bărbat, prinsu de jidovi, cînd vrea să-l omoară, sosiiu cu viteajii și-l scoș, înțelegînd că e **rimlean**.	N. TEST.	Noul Testament sau împăcarea au leagea noao... În cetatea Bălgradului, 1648.
s'au	1809	Bessarabia	Democratul Basarabiei	1575	Să-l gioare cum nu **s'au** tins la marha priatnicului său.	PALIA	Palia d'Orăştie. (1581–1582). I: Préface et livre de la Genèse, publiés avec le texte hongrois de Heltai et une introduction par Mario Roques. Les premières traductions roumaines de l'Ancien Testament. Paris, Champion, 1925.
săcurea	1950	Transylvania	Vatra romaneasca	1574	Amu și **săcurea** lângă rădăcinile lemnului zace.	CORESI, TETR.	Tetraevanghelul diaconului Coresi. Reimprimat după ediţia primă din 1560-61 de arhiereul dr. Gherasim Timuş Piteşteanu. Cu o prefaţă de Constantin Erbiceanu. Bucureşti, Tipografia Cărţilor Bisericeşti, 1889.

săl	1829	Wallachia	Curier romanesc	1591	Ca să-i fie lui direaptă ocină . . . , lui și cuconilor **săl** și a tot rodul său	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
suntu	1876	Bessarabia	Curierul	1521	„Cin[r]e eşti, D[oam]ne?“ Elu dzise: Eu **suntu** I[sus] cela ce tu-l goneşti!	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
temeiată	1893	Moldavia	Convorbiri literare	1648	Suflară vînture și loviră în casa aceaia și nu căzu, că era **temeiată** pre piatră.	N. TEST.	Noul Testament sau împăcarea au leagea noao... În cetatea Bălgradului, 1648.
tememu	1890	Transylvania	Aradu	1500	Nu ne **tememu** cîndu se va smenti pămîntul.	PSALT. HUR.	Psaltirea Hurmuzaki. [cca. 1500–1510] [Manuscris de la începutul secolului al XVI-lea; Biblioteca Academiei Române, cota 3077].
trăesc	1968	Bessarabia	Rev. Univ. Chisinau	1671	Făcutu-ți-um izvod întiiași dată de mari și vestiți istorici, cărora **trăesc** și acmu scrisorile în lume și vor trăi în veci.	M. COSTIN	Miron Costin, Opere. Ediţie critică cu un studiu introductiv, note, comentarii, variante, indice şi glosar de P. P. Panaitescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1958.
tuturoru	1865	Wallachia	Telegraful roman	1574	Cela ce va toată leagea feri și va greși ura fi[-va vi]novatu **tuturoru**. COD. VOR	CORESI, EV.	Diaconul Coresi, Carte cu învăţătură (1581), publicată de Sextil Puşcariu şi Alexie Procopovici. Vol. I. Textul. Bucureşti, Atelierele Grafice Socec et Comp., 1914. (Comisia Istorică a României).
unchiu	1950	Transylvania	Vatra romanească	1621	Iară vara premare sau vărul premare al maichei meale este-mi **unchiu** mic și leale mică.	PRAV. GOV.	Pravila bisericească numită cea mică, tipărită mai întâi la 1640 în Mănăstirea Govora; publicată acum în transcripţiune cu litere latine de Academia Română [Editată de A. I. Odobescu]. Bucureşti, Tipografia Academiei Române (Laboratorii români), 1884.

uritu	1870	Moldavia	Convorbiri literare	1600	Vai de omul cela de doce prescurea la besearecă cu **uritu**	CUV. D. BĂTR.	B. Petriceicu-Hasdeu, Cuvente den bătrâni. Tomul I. Limba română vorbită între 1550–1600. Studiu paleografico-linguistic de B. Petriceicu Hasdeu...; Tomul II. Cărţile poporane ale românilor în secolul XVI … ; Tomul III. Istoria limbei române. Partea I. Principie de linguistică. Ediţie îngrijită, studiu introductiv şi note de G. Mihăilă. Bucureşti, Editura Didactică şi Pedagogică, [vol.] I–III, 1983-1987.
urulu	1865	Wallachia	Curierul romanesc	1521	Argint sau aur sau veşmente nece de la **urulu** n-am jeluit.	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
vaiera	1870	Transylvania	Albina	1693	Foamea dară începînd, mănuntaiele i se întorcea şi stomahul tare îl chinuia... Vai maţele!... Vai pîntecele! Vai vintrile mele! a se vaieta şi a se **vaiera** începu.	CANTEMIR, IST.	Operele principelui Dimitrie Cantemir. Tipărite de Societatea Academică Română. Tom.VI: Istoria ieroglifică. (Operă originală inedită scrisă în limba românească la 1704). Bucureşti, 1883. [Editată de A. I. Odobescu].
vare	1893	Moldavia	Convorbiri literare	1646	Cine va ocoli locul acela.. de **vare** fi cu ştirea vemeşilor... tot negoţul ca să-ş piarză.	PRAV. MOLD.	Carte românească de învăţătură de la pravilele împărăteşti şi de la alte giudeaţe, cu zisa şi cu toată cheltuiala lui Vasilie Voivodul. Iaşi, l646.
voru	1876	Bessarabia	Curierul	1521	Aşa grăiţi şi aşa faceţi, cum cu leagea slobodziloru celora ce **voru** giudeţu se preemească.	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
vostră	1878	Wallachia	Pressa	1600	Cum mînie mea să nu să aprinză ca focul şi să nu arză, aşe ca să nu o paţă nime stînge păntru răotate faptelor vostră	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).

zâua	1913	Bessarabia	Cuvant moldovenesc	1793	Vizita de luare de **zâua** bună.	NARCHIS	J. J. Rousseau, Narchis sau îndrăgitul însuşi de sine. Comedie de... tălmăcită de I[oan] C[antacuzino] la anu [1]794. [Copie manuscrisă din 1803; Biblioteca Academiei Române, cota 3099].
zîc	1949	Transylvania	Vatra romaneasca	1675	Au pecinginat cu rapură de ceea ce-i **zîc** elefantia... de mînca într-o şezută cîte trei modii de pită.	DOSOFTEI, V. S.	Mitropolitul Dosoftei, Viaţa şi petreacerea svinţilor. Iaşi, 1682.

Appendix 3 Recovering old lema according with eDTLR (a sample)

Our corpus				eDTLR				
NotInDict	**Year**	**Region**	**Article**	**Year**	**Quotations**	**Lemma**	**Sigla**	**Title**
dechiara	1856	Transylvania	Federațiunea	1868	Acel sistem [ortografic] este dator sau a învinge noua oposițiune prin arguminte mai tari sau a se **dechiara** pe sine de învins.	DECLARA	CONV. LIT.	Convorbiri literare. Iaşi, apoi Bucureşti. Anul I (1867–1868) ş. u.
decisiunea	1878	Wallachia	Pressa	1939	Dl. Sextil Puşcariu, dimpotrivă, exprimă cu toată **decisiunea** punctul de vedere creştin.	DECIZĂTOR	STĂNILOAE, O.	Dumitru Stăniloae, Ortodoxie şi românism. Sibiu, Tiparul Tipografiei Arhidiecezane, 1939.
deplinirei	1829	Moldavia	Albina romanească	1829	De datorie este a nu mai întârzie de a aduce limba vorbită de mai mulți decât patru milioane români la gradul **deplinirei**	DEPLINIRE	PLR	Presa literară românească. Articole-program de ziare şi reviste (1789–1948). Ediție în două volume, note, bibliografie şi indici de I. Hangiu. [Bucureşti], Editura pentru Literatură, 1968.
desvălit	1918	Bessarabia	Sfatul țării	1806	De unde **desvălit** se veade că Vopiscus au fost om ca acela, căruia nu-i era greață a minți.	DEZVELIT	MAIOR, IST.	Petru Maior de Dicio-Sân-Martin, Istoria pentru începutul românilor în Dachia.Întocmită de..., iar acum pretipărită prin Iordachi de Mălinescu... Buda, Tipariul Crăieştei Tipografie a Universitatei Ungureşti, 1834.

elementu	1875	Transylvania	Telegrafulu romanu	1932	Acesta este **elementu** lui, fără de care nu poate el să răsufle.	ELEMENT		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
emancipărei	1878	Wallachia	Pressa	1861	Daca este vreo chestiune arzătoare pentru țară, este aceea a **emancipărei** țăranilor.	EMANCIPARE	PONTBRIANT, D.	Raoul de Pontbriant, Dicţionar româno-frances. Bucureşti – Göttingen – Paris, 1862.
esperimente	1880	Moldavia	1 Curierul. Foaia intereselor generale	1881	Țara, prin tristele **esperimente** la care e supusă de domnia frazei, va ajunge până în sfârşit să fie conservatoare.	EXPERIMENT	EMINESCU, O.	Mihai Eminescu, Opere. Ediţie critică îngrijită de Perpessicius. [Vol.] VI. Bucureşti, Editura Academiei, 1963; [vol.] VII ş.u. [Ediţie critică întemeiată de Perpessicius, îngrijită de un colectiv de cercetători de la Muzeul Literaturii Române], Bucureşti, Editura Academiei, 1977 ş.u. De la vol IX – sub titlul Publicistică, vol. I ş.u., 1980 ş.u.
iconom	1925	Bessarabia	Luminatorul	1982	**Iconom** să înțelege cel ce, cu puțin, din vreme să îngrijaște.	ECONOM	ZANNE, P.	Iuliu A. Zanne, Proverbele românilor din România, Basarabia, Bucovina, Ungaria, Istria şi Macedonia. Vol. I–X. Ediţie îngrijită de Mugur Vasiliu. Bucureşti, Asociaţia Română pentru Cultură şi Ortodoxie, Editura Scara, 2003–2004.
legati	1847	Transylvania	Organulu Luminarei	1582	Limbile... lor era **legati** în parî de foc și nu pute grăi.	LEGAT	COD. TOD.	Dr. Nicolae Drăganu, Două manuscripte vechi. Codicele Todorescu şi Codicele Marţian. [secolul XVII-lea] Studiu şi transcriere de... Ediţiunea Academiei Române. Bucureşti, Librăriile Socec et. Comp. şi C. Sfetea; Leipzig, Otto Harrassowitz; Wiena, Gerold et Comp., 1914.

legiuitore	1878	Wallachia	Romanulu	1874	A domnit... armonie între guvern şi puterea **legiuitore**.	LEGIUITOR	MAIORESCU, D.	Titu Maiorescu, Discursuri parlamentare, cu priviri asupra dezvoltării politice a României... Vol. I: (1866–1876); vol. II: (1876–188l); vol. III: (1881–1888); vol. IV: (1888–1895). Bucureşti, Socec, 1897, 1899, 1904; vol. V: (1895–1899). Bucureşti, Editura Minerva, 1915.
lejea	1913	Bessarabia	Cuvant moldovenesc	1960	M-am lăsat în **lejea** lui. ŞANDRU AMP;#8722; BRÎNZEU, JINA, 100. Cum îi în legea dumnealui: nu se merită să dai banii pe casa asta.	LEGE	MAT. DIALECT.	Materiale şi cercetări dialectale. [Vol.] I. [Bucureşti], 1960. (Academia Română. Filiala Cluj. Institutul de Lingvistică).
monăstirea	1893	Moldavia	Convorbiri literare	1851	**Monăstirea** monahilor franciscani plină de oameni.	MĂNĂSTIRE	BARIȚIU, P. A.	
neagiunsuri	1832	Moldavia	Albina romanească	1866	O adevărată critică, bazată pe minte matură şi pe cunoştinţe solide, ar fi descoperit destule **neagiunsuri** în prima ediţiune a acestu volum.	NEAJUNS	HASDEU, I. C.	B. Hasdeu, Istoria critică a românilor. Ediţie îngrijită şi studiu introductiv de Grigore Brâncuş. Studiu introductiv şi note de Manole Neagoe. Bucureşti, Editura Minerva, 1984.
nedomirire	1877	Wallachia	Romanulu	1881	Ajungînd de marginea eleşteului, se aşează şi el acolo jos şi privind cu **nedomirire**, ia aşa numai ca să zică şi el că face ceva, bălăcea cu nuiaua prin apă.	NEDUMERIRE	ISPIRESCU, L.	P. Ispirescu, Legende sau basmele românilor. Adunate din gura poporului. Bucureşti, Tipografia Academiei Române, 1892.
noptat	1949	Transylvania	Vatra romanească	1952	S-o **noptat**.	ÎNNOPTAT	ALR SN	Atlasul lingvistic român, serie nouă. Sub direcţia acad. Emil Petrovici. Redactor principal: Ion Pătruţ. Vol. I: A. Agricultură; B. Morărit; C. Grădinărit; D. Pomărit; E. Viticultură; F. Cânepa; G. Albinărit. Vol. II: A. Creşterea vitelor; B. Carul, căruţa,

								sania; C. Păsări de curte; D. Păstorit; E. Lâna, torsul, ţesutul; F. Meserii; G. Pădurărit. Vol. III: A. Plante; B. Câinele, pisica; C. Animale sălbatice; D. Păsări sălbatice; E. Vânătoare; F. Târâtoare, amfibii; G. Pescuit; H. Insecte; I. Timpul; J. Configuraţia terenului; K. Ape, umiditate, navigaţie; L. Comunicaţii; M. Naţiuni, categorii sociale, administraţie. [Bucureşti], Editura Academiei, vol. I–II: 1956; vol. III: 1961; vol. IV ş.u.: 1965 ş.u.12 (Academia Română. Institutul de Lingvistică şi Istorie Literară. Filiala din Cluj-Napoca).
pervi	1819	Bessarabia	Democratul Basarabiei	1900	Să-mi dai făină **pervi**.	PERVĂ	SĂGHINESCU, S.	V. Săghinescu, Scrutare Dicţionarului universal al limbei române, de Lazăr Şăineanu... Iaşi, 1898.
placu	1880	Transylvania	Gura Satului	1691	Au socotești cu mintea ta că-m' **placu** fap-rmarian57976 tel(e) tal(e) ceale făr de leage?	PLĂCEA		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
positive	1880	Wallachia	Pressa	1841	Cele de pre urmă . . . adecă pizma și zavistia au și **positive** și aducătoare de moarte însușiri.	POZITIV		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
precisiune	1893	Moldavia	Convorbiri literare	1874	Prescrie dar a stabili între idei și cuvinte un raport de **precisiune**.	PRECIZIE	MAIORESCU, CR.	Titu Maiorescu, Critice (1867–1892). Ediţie nouă şi sporită. Vol. I–III. Bucureşti, Socec & Comp., 1892–1893.

preposit	1890	Bessarabia	Candela	1806	Cel bătrîn... e acum în Blaj **preposit** Capitalului.	PREPOZIT	MAIOR, I. B.	Petru Maior, Istoria besericei românilor, atât a cestor dincoace precum şi a celor din colò de Dunăre. Buda, Crăiasca Tipografie a Universitatei din Peşta, 1813.
raportuiască	1893	Moldavia	Convorbiri literare	1833	Să **raportuiască** îndată vămii centrale din Eşi	RAPORTUI	URICARIUL	Theodor Codrescu, Uricariul cuprinzetoriu de hrisoave, anaforale şi alte acte din suta a XV– XIX, atingătoare de Moldova. Sub redacţiunea dsale... Vol. I–XXV. Iaşi, 1852–1895; pentru vol. I, II, III, v. şi ediţiunea a 2-a, Iaşi, 1871–1893.
răsboiului	1877	Wallachia	Pressa	1982	Isprava **răsboiului**, a cerni o lume.	RĂZBOI	ZANNE, P.	Iuliu A. Zanne, Proverbele românilor din România, Basarabia, Bucovina, Ungaria, Istria şi Macedonia. Vol. I–X. Ediţie îngrijită de Mugur Vasiliu. Bucureşti, Asociaţia Română pentru Cultură şi Ortodoxie, Editura Scara, 2003–2004.
răzemată	1950	Transylvania	Vatra romanească	1905	Fata şedea, sărmana, **răzemată** de doagele fîntînei.	REZEMAT	MERA, L. B.	I. T. Mera, Din lumea basmelor. Cu 87 ilustraţiuni originale. Bucureşti, „Minerva”, 1906.
rezolvirea	1919	Bessarabia	Democratul Basarabiei	1854	**Rezolvirea** ecuației simple cu o necunoscută.	REZOLVARE	LAZARINI, M.	Prof. G. Lazarini, Matematică. Partea I-a. Iaşi, Tipografia Româno-Francesă, 1854.
rostirei	1893	Moldavia	Convorbiri literare	1838	Sau pe amvona **rostirei**, prin ritorice cuvinte Învingînd pe nedreptate, Demosten s-ar fi făcut Şi a văduvei sărmane ar fi razăm şi părinte.	ROSTIRE	ASACHI, S. L.	Gh. Asachi, Scrieri literare. [Vol. I şi II]. Ediţie îngrijită, cu prefaţă, note şi glosar de N. A. Ursu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, [1957]. (Biblioteca pentru toţi).

sbirii	1877	Transylvania	Telegraful roman	1872	Poporul a fost maltratat pe stradă de **sbirii** agiei.	ZBIR	ALECSANDRI, T.	Vasile Alecsandri, Opere complete. Partea întâia: Teatru. Volumul I: Canţonete comice, scenete şi operete; volumul II: Vodeviluri; volumul III: Comedii; volumul IV: Drame. Bucureşti, Editura Librăriei Socec & Comp., 1875.
scola	1865	Wallachia	Telegraful roman	1657	Ce domno să va **scola** întăi, acela va peri întâi	SCULA	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).
severşit	1896	Bessarabia	Desteptarea	1693	Voia aceasta a lor spre ce **severşit**? Spre alcătuire a doae firi într-una.	SĂVÎRȘI	CANTEMIR, IST.	Operele principelui Dimitrie Cantemir. Tipărite de Societatea Academică Română. Tom.VI: Istoria ieroglifică. (Operă originală inedită scrisă în limba românească la 1704). Bucureşti, 1883. [Editată de A. I. Odobescu].
solenitate	1865	Wallachia	Telegraful roman	1841	În mijlocul cîntărilor bisericeşti şi a tămîierii, împăratul au intrat cu **solenitate** în Suceava.	SOLEMNITATE	FM	Foaie pentru minte, inimă şi literatură. Braşov, 1838–1865.
sorori	1868	Transylvania	Federatiunea	1854	A avut şi două **sorori**.	SORĂ	ARISTIA, PLUT.	C. Aristia, Plutarh. Parallela sau vieţele bărbaţilor illustri. Tradus din eleneşte de... Tomul I. Bucureşti, Tip. Colegiului Naţional, 1857.
spaciul	1868	Moldavia	Convorbiri literare	1819	Lipsindu-i **spaciul** de a da aice toate elementele organizării sale, să aruncăm deodată o repede privire asupra diferitelor medii.	SPAȚIU	CALENDAR	Calendariu pe anul de la Hristos 18... În Buda, 1817 ş. u.

straiu	1936	Bessarabia	Luminatorul	1982	Cine se-ntinde mai mult decît îi **straiu** îi rămîn picioarele goale.	STRAI		Studii şi cercetări lingvistice. [Bucureşti], Editura Academiei. Institutul de Lingvistică din Bucureşti. Anul I (1950) ş.u.
tesauru	1865	Transylvania	Foaia pentru Minte, Anima si Literatura	1860	**Tesauru** de monumente istorice pentru România, atît din vechi tipărite, cît și manuscripte	TEZAUR	TES.	A. Papiu Ilarian, Tesauru de monumente istorice pentru România, atât de vechi tipărite, cât şi manuscripte, cea mai mare parte străine, adunate, publicate cu prefaţiuni şi note ilustrate de... Tom. I–III. Bucureşti, Tipografia Naţională, 1862–1864.
theorie	1880	Wallachia	Pressa	1675	La bunătăţ şi facere şi **theorie** atîta s-au făcut, cît pre toți au întrecut.	TEORIE	DOSOFTEI, V. S.	Mitropolitul Dosoftei, Viaţa şi petreacerea svinţilor. Iaşi, 1682.
titula	1868	Moldavia	Convorbiri literare	1810	ca să să tipărească o carte ce poartă în frunte **titula**: „Cărticica năravurilor bune“	TITLU	IORGA, S. D.	N. Iorga, Studii şi documente cu privire la istoria românilor. Vol. I–XXXI. Bucureşti, Editura „Minerva”, 1901–1916.
ulcioraș	1970	Bessarabia	Povara bunatatii noastre	1913	A văzut biata băbușcă ce nu mai văzuse. Zestre nu-i, **ulcioraș** cu galbeni nu-i - calicie și măritiș, cât lumea!	URCIORAŞ	FURTUNĂ, V.	Dumitru Furtună, Vremuri înţelepte. Povestiri şi legende româneşti. Culese de… Bucureşti, Librăriile Socec & Comp., C. Sfetea şi Librăria Naţională, 1913. (Academia Română. Din viaţa poporului român. Culegeri şi studii. XV).
unicu	1890	Transylvania	Aradu	1908	Cân să făcea, să făcea mere şi nuci; cireș câte **unicu** singur.	UNIC	GRAIUL	Graiul nostru. Texte din toate părţile locuite de români, publicate de I. A. Candrea, Ov. Densusianu, Th. D. Speranţia. [Vol.] I: 1906; [vol.] II: 1908. Bucureşti, Socec & Co.

uritu	1870	Moldavia	Convorbiri literare	1600	Vai de omul cela de doce prescurea la besearecă cu **uritu**	URÂT	CUV. D. BĂTR.	B. Petriceicu-Hasdeu, Cuvente den bătrâni. Tomul I. Limba română vorbită între 1550–1600. Studiu paleografico-linguistic de B. Petriceicu Hasdeu...; Tomul II. Cărţile poporane ale românilor în secolul XVI … ; Tomul III. Istoria limbei române. Partea I. Principie de linguistică. Ediţie îngrijită, studiu introductiv şi note de G. Mihăilă. Bucureşti, Editura Didactică şi Pedagogică, [vol.] I–III, 1983-1987.
usul	1880	Wallachia	Romania libera	1851	Principele va avea vot și scaun în dieta Ungariei, în persoana sau prin delegații săi, după **usul** vechi.	UZ	BARIȚIU, P. A.	
vax	1893	Moldavia	Convorbiri literare	1939	Această carte e un **vax**.	VACS	SCRIBAN, D.	August Scriban, Dicţionaru limbii româneşti. (Etimologii, înţelesuri, exemple, citaţiuni, arhaizme, neologizme, provincializme). Ediţiunea întâia. Iaşi, Institutul de Arte Grafice „Presa Bună”, 1939.
vedé	1877	Transylvania	Telegraful roman	1574	Nedereptate nu putuiu **vedé** [n u o p u t u i u vedé V] și urii-o, e lege ta iubiiu. ib. 266. Potopeaște, Doamne, și împarți limbile lor Că văzui fărăleage și cearte în cetate.	VEDEA	CORESI, PS.	Coresi, Psaltirea slavo-română (1577) în comparaţie cu psaltirile coresiene din 1570 şi din 1589. Text stabilit, introducere şi indice de Stela Toma. [Bucureşti], Editura Academiei, 1976 (Academia Română. Comisia de texte româneşti vechi).
vedi	1880	Wallachia	Romania libera	1915	Făptura trupului era... așa precum să **vedi** arătat în gios	VEDEA	GCR	M. Gaster, Chrestomaţie română. Texte tipărite şi manuscrise (sec. XVI–XIX), dialectale şi populare, cu o introducere, gramatică şi un glosar româno-francez de… Vol. I: Introducere, gramatică, texte (1550–1710).

voru	1876	Bessarabia	Curierul	1521	Așa grăiți și așa faceți, cum cu leagea slobodziloru celora ce **voru** giudețu se preeme-ască.	VREA	COD. VOR.	Codicele voroneţean. Ediţie critică, studiu filologic şi studiu lingvistic de Mariana Costinescu. Bucureşti, Editura Minerva, 1981.
zidirei	1875	Transylvania	Telegrafulu romanu	1860	Se desfășur în ochi-mi minunile **zidirei**.	ZIDIRE	ALEXANDRESCU, O.	Gr. Alexandrescu, Opere. [Vol. I]. Ediţie critică, note, variante şi bibliografie de I. Fischer. Studiu introductiv de Silvian Iosifescu. [Bucureşti], Editura de Stat pentru Literatură şi Artă, 1957.
zilii	1880	Wallachia	Romania libera	1830	Domnul, dimpreună cu obicinuita obștească adunare … va hotărî plata **zilii** de lucru și a chiriii carălor.	ZI	REG. ORG.	[Regulamentul organic al Moldovei]. [Manuscris din 1831; Biblioteca Universităţii din Iaşi, cota VI – 61].

Appendix 4 Automatic correlations induced by 3 REPs & most frequent REPs in automatically correlated words from texts for each Romanian historical regions (a sample)

Regions	**Pair**	**REPs operations**	**Most frequent REPs in automatically correlated words**	**Percentage of most frequent REPs**
Moldavia	intămplare = întîmplară	ă->î e->ă i->î	u ->	11.55
	intămplatu = întîmplat	ă->î u-> i->î	s -> z	9.06
	esistinței = existenții	s->x ei->ii i->e	e -> ă	6.22
	inveățămentul = învățămîntul	e->ă e->î i->î	i -> î	5.82
	incependu = începînd	u-> e->î i->î	ei -> ii	3.58
	aședămentului = așezămîntul	ui-> d->z e->î	e -> i	3.09
	măngăemu = mîngîiem	ă->îi ă->î u->	i -> e	3.04
	invățămentului = învățămîntul	ui-> e->î i->î	a ->	2.39
	imbrățoșăndu = îmbrățișînd	oșă->işî u-> i->î	e -> î	2.34
	intrecendu = întrecînd	u-> e->î i->î	ă -> î	2.14
Wallachia	tranvayelor = tramvaiele	or->e n->m y->i	s -> z	12.89

	betrănului = bătrînul	ui-> ă->î e->ă	ei -> ii	6.27
	espoziţiunea = expoziția	s->z unea->a s->x	e -> ă	5.91
	imperatorele = împăratul	orele->ul e->ă i->î	ă -> î	3.69
	esportativei = exportatoarei	sx i->oa v->r	s -> x	3.58
	anecsiunilor = anexelor	cs->x i->e uni->	i -> e	3.22
	încunosciinţărei = înștiințării	cuno-> sc->ș ei->ii	e -> i	2.61
	îndealtmintrelea = altminterea	înde-> r-> l->r	u ->	2.61
	opoziţiunii = opoziției	s->z un-> ii->ei	ĭ -> i	2.04
	ocasiunea = ocazia	s->z un-> ea->a	a ->	1.97
Transylvania	serbâtoria = sărbătoarea	î->ă ria->area e->ă	u ->	18.26
	transsilbanieĭ = transilvaniei	ĭ->i b->v s->	s -> z	5.19
	essercitiului = exercițiul	ui-> t->ț ss->x	e -> ă	3.83
	romanescâ = românească	î->ă es->eas a->î	ei -> ii	2.56
	questoriului = chestorului	ui-> qu->ch i->	e -> i	2.10
	cercustantialminte = circumstantial	e->i us->ums minte->	i -> î	1.99
	cantaretie = cîntăreți	a->î tie->ți a->ă	a -> ă	1.97
	indrasnetiu = îndrăzneț	tiu->ț as->ăz i->î	ĭ -> i	1.77
	restempulu = răstimpul	e->ă e->i u ->	s -> x	1.48
	baserecesca = bisericeasca	es->eas e->i a->i	e -> î	1.38
Bessarabia	santementalu = sentimental	e->i a->e u->	ei -> ii	7.86
	ertăndu = iertînd	e->ie ă->î u->	u ->	5.79

	iantămplari = întîmplări	ia->î ă->î a->ă	s -> z	4.06
	ianțălegire = înțelegere	ia->î i->e ă->e	i-> ă	3.81
	domnievoastru = dumneavoastră	u->ă ie->ea o->u	i -> e	3.23
	combatarea = îmbătare	co->î a-> a->ă	a ->	2.73
	ianstreinatu = înstrăinat	ia->î e->ă u->	e ->ie	2.48
	iantămplari = întîmplări	ia->î ă->î a->ă	e ->i	2.07
	streinatățăi = străinătății	e->ă a->ă ă->i	ĭ -> i	1.99
	septămănei = săptămînii	e->ă ă->î e->i	i -> î	1.82

References

Ando, R. and Zhang, T. (2005). A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. JMLR.

Andriescu, A. (1977). Stil şi limbaj, Ed. Junimea, Iaşi.

Andriescu, A. (1979). Limba presei Româneşti în secolul al XIX-lea, Ed. Junimea, Iaşi.

Argamon-Engelson, S., Koppel, M. and Avneri, G. (1998). Style-based text categorization: What newspaper am I reading? In: Proceedings of the AAAI Workshop on Text Categorization, pp. 1-4.

Asher, N. and Lascarides, A. (2003). Logics of conversation. Cambridge: Cambridge University Press.

Avram, M., Sala, M., Vintila-Rădulescu, I. (coords). (1999). Institutul de Lingvistică „Iorgu Iordan“. 50 de ani de existență (1949-1999), Univers Enciclopedic, București.

Avramescu, A. & Avramescu, T (1996/2001). Pagini despre Basarabia, vol. I și II, Ed. Știința, Chișinău.

Baayen, R. H. (2008). Analyzing Linguistic Dat : A Practical Introduction to Statistics. Cambridge: Cambridge University Press.

Baldinger, K. (1952). Die Gestaltung des wissenschaftlichen Wörterbuchs. Historische Betrachtungen zum neuen Begriffssystem als Grundlage fur die Lexikographie von Hallig und Wartburg. In: Romanistisches Jahrbuch 5, pp. 65-94.

Baldinger, K. (1960). Alphabetisch order begrifflich gegliedertes Wörterbuchs? In: Zeitschrift für romanistische Philologie 76, pp. 521-53, [repr.in Zgusta ed. (1985), pp. 40-57].

Barbu-Mititelu, V. (2013) Rețea semantico-derivațională pentru limba română, București, Editura Muzeului Național al Literaturii Române.

Barbu-Mititelu, V., Dumitrescu, Ș, Tufiș, D. (2014). News about the Romanian Wordnet. In: Proceedings of the 7th International Global WordNet Conference, Tartu, Estonia.

Béjoint, H. (2010). The Lexicography of English. Oxford: Oxford University Press.

Bengio, Y, Simard, P., and Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. In: IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166.

Berger, A. L., Della Pietra, V. J., Della Pietra, S. A. (1996). A Maximum Entropy Approach to Natural Language Processing. In: Computational Linguistics, 22(1), pp. 39-71.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM: Review, 37, pp. 573-595.

Biber, D., & Conrad, S. (2009). Register, Genre, and Style. Cambridge: Cambridge University Press.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. In: Journal of Machine Learning Research, 3 (4-5), pp. 993-1022.

Blei, D. M., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning.

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. The Annals of Applied Statistics, pp. 17-35.

Blei, D. M. and Lafferty, J. D. (2009). Topic Models. In Text Mining: Classification, Clustering, and Applications, A. Srivastava and M. Sahami, Eds., ed London, UK: Chapman & Hall/CRC, pp. 71–93.

Blitzer, J. and Zhu, X. J. (2008). Semi-supervised Learning for Natural Language Processing. In: Tutorial Abstracts of ACL-08: HLT, Columbus, Ohio, USA, June 2008. Association for Computational Linguistics, p. 3.

Blum, A. and Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-training. COLT 1998.

Bobeică, A. (1993). Sfatul Țării. Stindard alrenașterii naționale, Chișinău, 1993.

Boian, E., Cojocaru, S., Ciubotaru, C., Colesnicov, A., Malahov, L., Petic, M. (2013). Language Technology and Resources for cultural and historic heritage digitization. In: Proceedings of the 2nd International Conference on Intelligent Information Systems 2013, August 20-23, 2013, Chișinău, Republic of Moldova, pp. 64-73.

Boian, E., Ciubotaru, C., Cojocaru, S., Colesnicov, A., Malahov, L. (2014). Cultural and Historical Heritage Digitization, Recognition and Conservation. In: Akademos: Revista de Ştiinţă, Inovare, Cultură şi Artă, nr. 1 (32), pp. 61-68.

Boldur, A. V. (1940). Contribuții la studiul Istoriei Românilor. Istoria Basarabiei, vol. III, Tiparul Moldovenesc, Chişinău.

Boldur, A. V. (1943). Basarabia românească, Bucureşti. Republ. in: Istoria Basarabiei,

vol. 2, Ed. Timpul, Chișinău, 2016.

Bratko, A, Cormack, G. V., Filipič, B., Lynam, T. R., Zupan, B. (2006). Spam Filtering Unssing Statistical Data Compression Models. In Journal of Machine Leanring Research 7, pp. 2673-2698.

Bréal, M. J. A. (1897). Essai de sémantique. Paris: Hachette.

Bréal, M. J. A. (1900). Semantics Studies in the Science of Meaning, transl. by Henry Cust, London: W. Heinemann.

Bresnan, J. (ed.) (1982). The Mental Representation of Grammatical Relations. Cambridge, MA: The MIT Press.

Burges, C. (1998). A Tutorial on SVM for Pattern Recognition, Data Mining and Knowledge Discovery, 2, pp. 121-167.

Carnap, R. (1972). Semnificaţie şi necesitate. Un studiu de semantică şi logică modală. Transl. by Gheorghe Enescu and Sorin Vieru, Ed. Dacia, Cluj-Napoca.

Carstensen, K-U., Ebert, C., Ebert, C., Jekat, S., Langer, H. and Klabunde, R. (eds.) (2009). Computerlinguistik und Sprachtechnologie: Eine Einführung. Spektrum Akademischer Verlag.

Carter, R. (1987). Vocabulary: Applied Linguistics Perspectives. London: Allen & Unwin.

Celikyilmaz, A., Hakkani-Tur, D., and Tur, G. (2010). Lda based similarity modeling for question answering. In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search, 1–9. Association for Computational Linguistics.

Chapelle, O., Schölkopf, B., Zien, A. (2006). Semi-Supervised Learning, The MIT Press, Cambridge, Massachusetts.

Charniak, E. (2000). A maximum-entropy - inspired parser. In: Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, Seattle, Washington, USA, pp. 132-139.

Chendi, I. (1900). Începuturile ziaristicii noastre (1789-1895), Orăștie, 1900;

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

Chomsky, N. (1981). Lectures on Government and Binding, Foris, Dordrecht.

Chomsky, N. (1995). Bare Phrase Structure. In: G. Webelhuth (ed.), Government and Binding Theory and the Minimalist Program. Oxford: Blackwell, pp. 383-439.

Ciobanu, A. and Dinu, L. (2014). An Etymological Approach to Cross-Language

Orthographic Similarity. Application on Romanian. In: Proceedings of EMNLP-2014, Oct. 25-29, 2014, Doha, Quatar, pp. 1047-1058.

Ciobanu, A. M., Dinu, L. P., Șulea, O.-M., Dinu, A., Niculae, V. (2013). In: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Sofia, Bulgaria, August 8 2013, Association for Computational Linguistics, pp. 102-106.

Ciobanu, Ș. (1933). Din istoria mișcării naţionale în Basarabia [II]. In Viaţa Basarabiei, nr. 1-2, anul II, pp. 3-13.

Ciobanu, Ș. (1992a). Publicaţiile periodice românești. In: Cultura românească în Basarabia sub stăpânirea rusă, Ed. Encilclopedică Gheorghe Asachi, Chișinău, pp. 223-243.

Ciobanu, Ș. (1992b). Studii rusești asupra trecutului și culturii românilor, în Cultura românească în Basarabia sub stăpânirea rusă, Ed. Enciclopedică Gheorghe Asachi, Chișinău, pp. 245-256.

Claridge, C. Historical Corpora. In A. Lüdeling, & M. Kytö (Eds.) (2008). Corpus Linguistics. An International Handbook, Volume 1. Berlin: De Gruyter, pp. 242–259.

Coates, A., Ng, A. Y., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In: Journal of Machine Learning Research - Proceedings Track, 15, pp. 215-223.

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

Coman, M. (1997). Manual de jurnalism, vol. I, II, Polirom, Iași.

Coman, M. (1999). Introducere în sistemul mass-media, Polirom, Iași.

Constantin, I. (1994). Basarabia sub ocupație sovietică, de la Stalin la Gorbaciov, Ed. Fiat Lux – București.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. Journal of Semantics, 12(1), pp. 15-68.

Corfis, I. A., O'Neill, J., and Beardsley, Jr. T. S. (eds.) (1997). Early Celestina Electronic Texts and Concordances. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Coşeriu, E. (1997). Sincronie, diacronie şi istorie. Problema schimbării lingvistice, versiune în limba română de Nicolae Saramandu, Ed. Enciclopedică, Bucureşti.

Coşeriu, E., (1966). Les théories linguistiques et leurs possibilités d'application. Structures lexicales et enseignement du vocabulaire. In: Actes du premier colloque

international de linguistique appliquée. Presse Universitaire de Nancy.

Coteanu, I., Seche, L., Seche, M. (coords.) *et al.* (1975/ 1996). Dicționarul explicativ al limbii române (DEX), București.

Coval, D. (1991). Ecouri dintr-un secol, Pagini din presa periodică moldovenească sec. XIX, Editura Universitas, Chișinău.

Coval, D. (1992). Din istoria jurnalisticii românești, Ed. Ştiinţa, Chişinău.

Cristea, D. & Tufiș, D. (2002). Resurse lingvistice românești şi tehnologii informatice aplicate limbii române. In: Ichim, O., Olariu, F.-T. (eds.), Identitatea limbii și literaturii române în perspectiva globalizării, Romanian Academy, Institute of Romanian Philology "A. Philippide", Trinitas Publishing House, Iași.

Cristea, D. and Tufiș, D. (1985). Quernal: A Natural Language Interface for Relational Data Bases. Simpozionul Național de Inteligență Artificială, ed. a II-a, București.

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

Cristea, D., Haja, G., Răschip, M, Moruz, A. M., Pătrașcu, M. (2011). Statistici parțiale la încheierea proiectului eDTLR – Dicționarul Tezaur al Limbii Române în format electronic. In volumul Lucrările conferinței naționale „Limba română: ipostaze ale variației lingvistice", 3-4 decembrie 2010, Catedra de Română, Universitatea din București.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). The Digital Form of the Thesaurus Dictionary of the Romanian Language. In: Proceedings of SpeD 2007 Speech Technology and Human - Computer Dialogue, Iași, May 10-12, 2007.

Cristea, D., Simionescu, D., Haja, G. (2012). Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations. In: Proceedings of LREC-2012, Istanbul, 21-25 May.

Croft, W. and Cruse, D. A. (2004). Cognitive linguistics. Cambridge: Cambridge University Press.

Cruse, D. A. (1986). Lexical Semantics, University of Cambridge Press.

Cruse, D. A., Hundsnurscher, F., Job, M. and Lutzeier, P. R. (eds.) (2002). Handbook of Lexicology, Vol. 1, Berlin & New York: Walter de Gruyter.

Cruse, D. A. (2000). Meaning in Language: An Introduction to Semantics and Pragmatics, Oxford: Oxford University Press.

Crystal, D. (1987). The Cambridge Encyclopedia of Language, Cambridge: Cambridge

University Press.

Curteanu, N, Moruz, A., Trandabăț, D. (2008) Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing in Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), Manchester, pp. 55.

Curteanu, N. (1983). Algoritmi de analiză sintactică a frazei și propoziției românești. In: Proceedings INFO-IAŞI'83, pp. 533-548.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. In: Audio, Speech, and Language Processing, IEEE Transactions on, 20(1), pp. 30–42.

Danilov, M. (2012). Presă și cenzură în Basarabia. Documentar secolul al XIX-lea-începutul secolului XX, Ed. Pontos, Chișinău.

Danilov, M. (resp.) (2005). Presa din Basarabia. Analize, contexte, valori (1854-2004), Cartdidact.

Danzin, A. (1992). Towards a European Language Infrastructure, March.

Dascălu, M. and Gîfu, D. (2015). Evaluating the Complexity of Online Romanian Press at the 11th International Conference Linguistic Resources and Tools for Processing the Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania, pp. 149-162.

Dascălu, M., Dessus, P., Trăușan-Matu, Ş., Bianco, M., Nardy, A. (2013a). ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. Lecture Notes in Computer Science (LNAI), vol. 7926, pp. 379–388. Springer, Heidelberg.

Dascălu, M., Trăușan-Matu, Ș, Dessus, P., Bianco, M., Nardy, A. (2013b). ReaderBench, o platformă integrată pentru analiza complexității textuale și a strategiilor de lectură. In: RoCHI, Ștefănuț, T & Rusu, C (eds.), pp. 39-49.

Davies, M. (2013). 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. pp. 46-67.

Davies, M. (2015). Corpus of Historical American English (COHA), Harvard Dataverse, V1.

Dănăilă, I., Michăescu, R. (1980). Banca de date fono-morfo-semantice a limbii române (BANDASEM). In: Limba Română, vol. 29, nr. 5, Bucureşti.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information

Science 41(6), pp. 391-407.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997). Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4), pp. 380–393.

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

Delmonte, R. (2007). Computational Linguistic Text Processing, vol. I&II, Ed. Nova Science Publishers, Inc., New York.

Densusianu, O. (1902). Filologia Romanică în universitatea noastră, Bucuresci, J. V. Socecu Editeur.

Densușianu, O. A. (1901/1938). Histoire de la langue roumaine. I, II. Paris: Ernest Leroux. Rom. edn.: Byck, J. (ed.) (1961). Istoria limbii române. I, II, Ed. Științifică, București.

Diaconescu, P. (1974). Elemente de istorie a limbii române literare moderne. Partea I. Probleme de normare a limbii române literare moderne (1830–1880), Bucureşti, pp. 5-6.

Dicționarul Limbii Române (2006) Tomul I, Partea a 4-a, Litera D, Deja-deținere, București, Editura Academiei Române.

Dinu, A. and Dinu, L. P. (2005). On the Syllabic Similarities of Romance Languages. In: Gelbuch, A. (ed.), CICLing, 2005, LNCS 3406, pp. 774-777.

Dinu, L. (2003). On the classification and aggregation of hierarchies with different constitutive elements. In: Fundamenta Informaticae, 55, 1, pp. 39-50.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, pp. 103-130.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, pp. 103–130.

Domonkos, E.N. (1967). Algoritm de traducere automată din limba engleză în limba română, Ed. Didactică şi Pedagogică, Bucureşti, 1967.

Drăgan, I. (1996). Paradigme ale comunicării în masă, Ed. Șansa, București.

Drăgănescu, M. and Burileanu, C. (eds.) (1986). Analiza și sinteza semnalului vocal, Ed. Academiei, București.

Duda, R. O. and Hart, P. E. (1973). Pattern classification and scene analysis, Wiley, New York.

Dumitrescu, Ș. D. (2015). Rowordnetlib – the First Api for the Romanian Wordnet. In Proceedings of the Romanian Academy, Series A, vol. 16: 1, 00, pp. 87-94.

Dumitrescu, Ș. D., Ion, R. and Ștefănescu, D., Boroș, T. and Tufiș, D. (2012). Romanian to English Automatic MT Experiments at IWSLT12. In: Proceedings of the International Workshop on Spoken Language Translation. Hong Kong, pp. 136-143.

Erjavec, T. (2012). The goo300k corpus of historical Slovene. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC'12, Paris, ELRA.

Evans, V. and Green, M. (2006). Cognitive Linguistics: An Introduction. Edinburgh: Edinburgh University Press.

Fellbaum, C. (ed.). (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Feng, S., Banerjee, R., and Choi, Y. (2012). Syntactic stylometry for deception detection. In: Proceedings of ACL 2012: Short Papers, pp. 171-175.

Fillmore, C. J. (1976). Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280(1), pp. 20-32.

Firth, J. R. (1957). Papers in linguistics 1934–51. Oxford University Press.

Firth, J. R. (1962). A Synopsis of Linguistic Theory, 1930-1955. In: Studies in Linguistic Analysis, Basil Blackwell, Oxford, p. 22.

Frank, E. and Hall, M. (2001). A simple approach to ordinal prediction. European Conference on Machine Learning. Freiburg, Germany. Springer-Verlag. 145-156.

Frege, G. (1892). On Sense and Meaning (first published in Zeitschrift fur Philosophie und Philosophische Kritik, Vol. 100, pp. 25-30, English trans. On Sense and Meaning. In: Geac – Black (eds.), 1952, pp. 56-78. Reprinted in Translations from the Philosophical Writings of Gottlob Frege (edited by P. Geach and M. Black), Oxford: Basil Blackwell, Oxford, 1952 (3rd Editions 1980).

Frinken, V., Zamora-Martinez, F., Espana-Boquera, S., Castro-Bleda, M.J., Fischer, A., Bunke, H. (2012). Long-short term memory neural networks language modeling for handwriting recognition, Pattern Recognition. In: 21st International Conference on Pattern Recognition (ICPR- 2012), pp. 701-704.

Fromkin, V., and Rodman, R. (1988). An Introduction to Language. 4th ed. New York: Harcourt Brace Jovanovich.

Frunză, O., Inkpen, D., and Nadeau, D. (2005). A text processing tool for the Romanian language. In: Proceedings of the EuroLAN 2005 Workshop on Cross-Language Knowledge Induction.

Gazdar, G., Klein, E. H., Pullum, G. K., Sag, I. A. (1985). Generalized Phrase Structure

Grammar. Cambridge, MA.

Geeraerts, D. (2010). Theories of Lexical Semantics: A Cognitive Perspectives. Oxford: Oxford University Press.

Ghibu, O. (1926). De la Basarabia rusească la Basarabia românească, Cluj-Napoca.

Gîfu, D (2016a). Diachronic Analysis Using a Statistical Model. In: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI'2016, July 25-29, Chișinău, Republic of Moldova.

Gîfu, D (2016d). The Analysis of Diachronic Variation in Romanian Print Press. In: Proceedings of the first PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD 2016, Carretero, J, Garcia Blas, J., Petcu, D. (eds.), Computer Architecture, Communications, and Systems Group (ARCOS), Spain, Feb. 2016, pp. 49-53.

Gîfu, D. (2015a). Creating of Parallel Lexicons for Romanian and Moldovan. In: Communication, Context, Interdisciplinarity, Iulian Boldea (coord.), vol. 3, Ed. "Petru Maior" University, Târgu-Mureş, 2015, section: Language and discourse, pp. 617-626.

Gîfu, D. (2015b). Contrastive Diachronic Study on Romanian Language. In: Proceedings of first Workshop on Foundation of Informatics, FOI-2015, Cojocaru, S. and Gaindric, C. (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, 2015, pp. 296-310.

Gîfu, D. (2015c). Particularities of Romanian Journalistic Language in the 19th century at the International Interdisciplinary Doctoral Conference, IIDC 2015, first edition, 25-26 Sept. 2015, Bucharest.

Gîfu, D. (2016b). Diachronic Evaluation of Newspapers Language between Different Idioms at the NLPMJ held at the International Joint Conference on Artificial Intelligence, IJCAI-16, 9-15 July 2016, New York, USA.

Gîfu, D. (2016c). The Chronology of Old Romanian Words at the 3rd International Conference Globalization, Intercultural Dialogue and National Identity, GIDNI, Language and Discourse section, 19-20 May 2016, Târgu-Mureș, Romania.

Gîfu, D. and Simionescu, R (2016). 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., Dascălu, M., Trăușan-Matu, Ș, Allen, L. (2016). Time Evolution of Writing Styles in Romanian Language at the 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016.

Goddard, C. (1998). Semantic Analysis: A Practical Introduction. Oxford: Oxford

University Press.

Goddard, C. (2009). The Natural Semantic Metalanguage Approach. The Oxford Handbook of Linguistic Analysis, Heine, B. and Narrog, H. (eds.).

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of English books. In: Proceedings of *SEM 2013, pp. 241-247.

Golub, G. H. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis 2(2), pp. 205-224.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. Numerische Mathematik 14(5), pp. 403-420.

Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H., Gonzalez-Rodriguez, J., Moreno, P. J. (2014). Automatic Language Identification using Long Short-Term Memory Recurrent Neural Networks. Interspeech, pp. 2155-2159.

Goodman, J., Heckerman, D., and Rounthwaite, R. (2005). Stopping spam. Scientific American, 292(4), pp. 42–88.

Gooskens, C. (2006). 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, pp. 101-113.

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

Gorunescu, F. (2011). Data Mining. Concepts and Techniques. Intelligent Systems Reference Library, vol. 12, Springer.

Graves, A. (2008). Supervised Sequence Labelling with Recurrent Neural Networks. Technische Universität München. http://deeplearning.cs.cmu.edu/pdfs/1104/Supervised_Sequence_Labeling.pdf

Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in Neural Information Processing Systems, pp. 545–552.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In: Proceedings of the 24th Annual Conference of the Cognitive Science Society.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In Neural information processing systems 15. Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the

National Academy of Science, 101, pp. 5228-5235.

Grossman, D. A. & Frieder, O. (2004). Information retrieval: algorithms and heuristics, 2nd ed., vol. 15 of the Information Retrieval Series, Springer.

Grossu, S. (2003). Presa din Basarabia în contextul sociocultural al anilor 1906-1944, Ed. Tehnica-Info, Chişinău.

Guiraud, P. (1958). La grammaire, Presses Universitaires de France, Paris, p. 76.

Guiraud, P. (1964). La sémantique (= Que sais-je? 655), Paris: PUF.

Gullberg, M., & Indefrey, P. (eds.). (2010). The earliest stages of language learning. Malden, MA: Wiley-Blackwell.

Gupta, V. and Lehal, G. S. (2009). A survey of text Mining Techniques and Applications in "Journal of Emerging Technologies in Web Intelligence", vol. I, no. I, pp. 60-61.

Haghighi, A. and Klein, D. (2006). Prototype-driven Learning for Sequence Models. HLT/NAACL 2006.

Haja, G., Dănilă, E., Forăscu, C., Aldea, B.-M. (2005). Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea, Ed. Alfa, Iași.

Hajič, J., Hric, J., and Kuboň, V. (2000). Machine translation of very close languages. In: Proceedings of the 6th Applied Natural Language Processing Conference, pp. 7-12. Association for Computational Linguistics., http://dx.doi.org/10.7910/DVN/8SRSYK.

Hangiu, I. (1987). Dictionar al presei literare românești (1790 – 1982), Ed. Științifică și Enciclopedică, București.

Hangiu, I. (2004). Dicționarul presei literare românești: (1790-2000), Ed. Institutului Cultural Român București, 2004.

Hanks, P. (ed.) (2007). Lexicology: Critical Concepts in Linguistics, Routledge.

Harris, Z. (1952). Discourse analysis. Language, 28:18-23.

Herrera, M. T. and González de Fauve, M. E. (1997). Concordancias Electrónicos del Corpus Médico Español. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Higginbotham, J. (1989). Elucidations of Meanung. In Linguistic and Philosophy 12, p. 470.

Hochreiter, S and Schmidhuber, J. (1997b). Long Short-Term Memory. In: Neural Computation 9(8), pp. 1735-1780.

Hochreiter, S. and Schmidhuber, J. (1995). Long short-term memory. Technical Report FKI-207- 95, Fakultät für Informatik, Technische Universität München.

Hochreiter, S. and Schmidhuber, J. (1996). Bridging long time lags by weight guessing

and “Long Short-Term Memory”. In Silva, F. L., Principe, J. C., and Almeida, L. B. (eds.) Spatiotemporal models in biological and articial systems. IOS Press, Amsterdam, Netherlands. Serie: Frontiers in Articial Intelligence and Applications, Volume 37, pp. pages 65-72.

Hochreiter, S. and Schmidhuber, J. (1997a). LSTM can solve hard long time lag problems. In: Advances in Neural Information Processing Systems 9. MIT Press, Cambridge MA. Presented at NIPS 96.

Hochreiter, S., Bengio, Y., Frasconi, P. and Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In: A Field Guide to Dynamical Recurrent Neural Networks, Kremer, S. C. and Kolen, J. F. (eds.), IEEE Press.

Hodos, N. & Sadi Ionescu, A. (resp.) (1913). ***Publicațiunile periodice românești (ziare, gazete, reviste), Tom I Catalog alfabetic (1820 - 1906), București, Leipzig, Viena.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning Journal, 42(1), pp. 177-196.

Hristea, T. (1981). Sinteze de limba română, the second edition, Ed. Didactică și Pedagogică, București, pp. 100-102.

Hristea, T. (1981). Principiile ortografiei româneşti actuale. In: Hristea, T. (coord.), Sinteze de limba română, ediţia a doua, revăzută şi mult îmbogăţită, Editura Didactică şi Pedagogică, Bucureşti, 1981, pp. 99-109.

Hudson, R. (1986/1995). Identifying the linguistic foundations for lexical research and dictionary design. In: D. Walker, A: Zampolli and N. Calzolari (eds.) Automating the Lexicon: Research and practice in a Multilingual Enviroment. Oxford: Oxford University Press, pp. 21-52.

Hurford, J. and Heasley, B. (1983). Semantics: A coursebook. Cambridge: Cambridge University Press.

Iftene, A. and Balahur-Dobrescu, A. (2008). Named Entity Relation Mining Using Wikipedia. In Proceedings of the Sixth Language Resources and Evaluation Conference, Marrakech, Morocco.

Iliescu, I. (1999). O istorie deschisă a presei române (1565-1925), Ed. Mirton, Timișoara.

Ilincan, V. (2010). Istoria presei literare românești. Secolul al IXI-lea, Suceava.

Ion, R. and Tufiş, D. (2004). Multilungual Word Sense Disambiguation Using Aligned

Wordnets. In Romanian Journal of Information Science and Technology, vol. 7, no 1 - 2, pp. 183-200.

Ion, R. and Tufiş, D. (2009). Multilingual versus Monolingual Word Sense Disambiguation. International Journal of Speech Technology, 12 (2-3), pp. 113-124.

Ion, R. şi Ştefănescu, D. (2011). Unsupervised Word Sense Disambiguation with Lexical Chains and Graph-Based Context Formalization. In Z. Vetulani, ed. 2011. LTC 2009. Heidelberg: Springer, pp. 435-443.

Ion, R., Ştefănescu, D., Ceauşu, A., and Tufiş, D. (2008). RACAI's QA System at the Romanian-Romanian Multiple Language Question Answering (QA@CLEF2008) Main Task. Carol Peters *et al.* (eds.) Working Notes for the CLEF 2008 Workshop: 10.

Ioncioaia, F. (2000). Introducere în presa scrisă. Ed. Universității "Alexandru Ioan Cuza", Iași.

Jackendoff, R. (1987). Consciousness and the computational mind. Cambridge, MA: MIT Press.

Jackendoff, R. (1990). Semantic structures. Cambridge, MA: MIT Press.

Jackendoff, R. (2011). Conceptual Semantics. In: Claudia Maienborn, Klaus von Heusinger, and Paul Portner (eds.), in Semantics: An International Handbook of Natural Language Meaning, Vol. 1, De Gruyter Mouton, pp. 688-709.

Jackendoff, Ray. (1983). Semantics and cognition. Cambridge, MA: MIT Press.

Jackendoff, Ray. (2002). Foundations of language. Oxford: Oxford University Press.

Jackson, H. (1988). Words and their meaning, London-New-York: Longman.

Jackson, H. and Zé Amvela, E. (2000). Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology.

Joachims, T. (1997). Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund.

Joachims, T. (1999). Making large-scale support vector machine learning practical. În Advances in kernel methods, MIT Press, pp. 169-184.

Johnson, M. (1987). The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason, Chicago: University of Chicago Press.

Juan A., Vilar D., Ney H. (2007). Bridging the gap between Naïve Bayes and Maximum Entropy. In: Proceedings of the PRIS 2007, Funchal (Portugal), pp. 59-65.

Jurafsky, D. and Martin, J. H. (2000). Speech and Language Processing, Prentice Hall, NJ, USA.

Jurafsky, D. and Martin, J. H. (2009) Speech and Language Processing. Prentice Hall,

2nd edition.

Kao, A. and Poteet, S. (2004). Text Mining and Natural Language Processing-Introduction for the Special Issue in SIGKDD Explorations, vol. VII, pp. 1-2.

Kasten, L., Nitti, J., and Jonxis-Henkemens, W. (1997). The Electronic Texts and Concordances of the Prose Works of Alfonso X, El Sabio. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Katz, J. J. and Fodor, J. A. (1963). The Structure of a Semantic Theory. In: Language, 39, pp. 170-210.

Kempson, R. M. (1977). SemanticTheory. Cambridge: Cambridge University Press.

Kenter, T., Erjavec, T., Dulmin, M. Z., Fišer, D. (2012). Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In: Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Avignon, France, 24 April 2012, pp. 1-6.

Klavans, J. (Ed.). (1995). Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity. AAAI Press, Menlo Park, CA. AAAI Technical Report SS-95-01.

Klein, D and Manning, C. (2003). Maxent model, conditional estimation, and optimization. ACL 2003 tutorial.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11), Portland, Oregon.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). Dirichlet and Inverted Dirichlet Distributions. In: Continuous Multivariate Distributions. vol. 1: Models and Applications, ed New York, NY: Wiley, pp. 485–527.

Kulikovski, L. (1957). Presa basarabeană de la începuturi pînă în anul 1957. Catalog, Biblioteca Municipală B. P. Hasdeu.

Kulikovski, L. (coord.) (2002). Presa basarabeană de la începuturi pînă în 1957: catalog / Bibl. Municipală "B. P. Hasdeu"; Bibl. „M. Lomonosov"; alcăt.: M. Şcelcikova, Ed. Epigraf, Chişinău.

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

Kytö, M., & Pahta, P. (2012). 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, pp. 123-133.

Lafferty, J. D. and Blei, D. M. (2005). Correlated topic models. In: NIPS.

Lakoff, G. (1987). Women, Fire and Dangerous Things: What Categories Reveal About the Mind. Chicago: University of Chicago Press.

Landauer, T. K., Dumais, S.T. (1997): A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review 104(2), pp. 211–240.

Landauer, T. K., Foltz, P.W., Laham, D. (1998a). An introduction to Latent Semantic Analysis. Discourse Processes 25(2/3), pp. 259–284.

Landauer, T. K., Laham, D., Foltz, P. W. (1998b) Learning human-like knowledge by singular value decomposition: a progress report. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems, vol. 10, MIT Press, Cambridge, pp. 45-51.

Landgacker, R. W. (2000). Grammar and Conceptualization, Mouton de Gruyter, Berlin.

Lascarides, A. and Asher, N. (2007). Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure. In: H. Bunt and R. Muskens (eds.) Computing Meaning: Volume 3, Springer, pp. 87-124.

Leech, G., Hundt, M., Mair, C. and Smith, N. (2009). Change in Contemporary English: A Grammatical Study. Cambridge: Cambridge University Press.

Lehrer A. (1974). Semantic fields and lexical structure. Amsterdam: North Holland.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals., Doklady Akademii Nauk SSSR 163(4), pp. 845-848.

Levit, E. (1979). Primele încercări de editare a unui periodic în Basarabia şi răspândirea iniţială a presei ruse în ţinut (1820-1850). In: Pagini din istoria literaturii şi culturii moldoveneşti, Ştiinţa, Chişinău.

Lewis, D. (1998). Naïve Bayes at Forty: The Independence Assumption in Information Retrieval. In: Proceedings of ECML-98, pp. 4-15.

Li, Y and Chung, S. M. (2005). Text document clustering based on frequent word sequences. In: International Conference on Information and Knowledge Management, CIKM '05, Bremen, Germany, pp. 293-294.

Löbner, S. (2002). Understanding Semantics. London: Arnold. Longman.

Lu, Y., Okada, S., Nitta, K. (2013). Semi-supervised Latent Dirichlet Allocation for Multi-label Text Classification. In: Proceedings of the 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE

2013, Amsterdam, The Netherlands, June 17-21, vol. 7905, Springer Berlin Heidelberg, pp. 351-360.

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

Lupu, I., Camariano, N. and Papadima, O. (resp.) (1966/1972). ***Bibliografia analitică a periodicelor românești (1780 – 1850), vol. I(1-3), vol. II(1-3), București, 1966, 1972.

Lyons, J. (1977). Semantics 1 & 2. Cambridge, CUP. [one of the "Bibles" of semantics; it is old but complete. Quite dense, use only for reference of specific points].

Maior, P. (1819). Orthografia Romana sive Latino-Valachica, Buda.

Malakasiotis, P. (2010). AUEB at TAC 2009. In: Proceedings of the Second Text Analysis Conference (TAC 2009) November 16-17, 2009 National Institute of Standards and Technology Gaithersburg, Maryland, USA.

Manning, C. and Schütze, H. (1999). Foundations of Statistical Natural Language Proceesing, MIT Press. Cambridge, MA.

Marcus, S. (1970b). Poetica matematică, Ed. Academiei, Bucureşti.

Marcus, S. (ed.) (1970a). Metode distribuţionale algebrice în lingvistică, Ed. Academiei, Bucureşti.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

Marian, L. (1932). Începuturile publicisticii românești. In Viaţa Basarabiei, nr. 2.

Martin, J. H. (1996). Computational approaches to figurative language. Metaphor and Symbolic Activity, 11(1), pp. 85-100.

Mâţu, C. (1930). O necesitate desconsiderată: presa românească în Basarabia, Chişinău.

McCulloch, W. S. and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nevous Activity. In Bulletin of Mathematical Biophysics, vol. 5, pp. 115-133.

Meillet, A. (1975). Linguistique historique et linguistique générale. Paris: Champion (reprint 1982).

Melamed, I. D. (1995). Automatic Evaluation and Uniform Filter Cascades for Inducing Nbest Translation Lexicons. In: Proceedings of the 3rd Workshop on Very Large Corpora, Boston, Massachusetts.

Mic dicționar academic, Ed. Univers Enciclopedic, 2001, 2002, 2003.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K. (2011). The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A.

Nowak, and E. Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. Science.

Mihail P. & Mihail Z. (1993). Acte în limba română tipărite în Basarabia, I, (1812-1830), precedate de Bibliografia tipăriturilor românești din Basarabia 1812-1830, București.

Mihalcea, R. and Năstase, V. (2012). Word epoch disambiguation: Finding how words change over time. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 259–263, Jeju Island, Korea, July. Association for Computational Linguistics.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus based and Knowledge-based Measures of Text Semantic Similarity. In: AAAI Conference on Artificial Intelligence.

Mikolov, T., Kombrink, S., Burget, L., Cernocky, J. H., Khudanpur, S. (2015). Learning Longer Memory in Recurrent Neural Networks. In: ICLR-2015 Workshop.

Milea, L. (1982). SIRLIN: Un sistem de întrebare-răspuns în limba română. In: Buletinul Român de Informatică, Bucureşti.

Miller, G. A. (1985). Wordnet: A Dictionary Browser. In: Information in Data, Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary. Waterloo, Canada: University of Waterloo.

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM 38(11), pp. 39-41.

Miller, G. A. (1998). Foreword. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, MIT Press, Cambridge, pp. xv–xxii.

Miller, G. A. (2010). WordNet. Princeton University Press, Princeton.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue), 3(4), pp. 235-312.

Moruz, M., Iftene, A., Moruz, A., Cristea, D. (2012). Semi-automatic alignment of old Romanian words using lexicons. In: Proceedings of the 8th International Conference „Linguistic resources and tools for processing of the Romanian language", Ed. Universității „Alexandru Ioan Cuza" din Iași, pp. 119-125. OCR (Optical Character Recognition) Technology http://www.unescap.org/stat/pop-it/pop-guide/capture_ch01.pdf

Mozer, M. C. (1989). A focused back-propagation algorithm for temporal sequence recognition. Complex Systems, no. 3 , pp. 349-381.

Mueller, J. and Thyagarajan, A. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In: Association for the Advancement of Artificial Intelligence

(http://www.mit.edu/~jonasm/info/MuellerThyagarajan_AAAI16.pdf).

Munteanu, E. (2008). Lexicologie biblică românească. Humanitas, București.

Mușat, C., Rizoiu, M.-A., Trăușan-Matu, Ș. (2010). O metodă pentru evaluarea și îmbunătățirea performanțelor sistemelor de extragere a topicelor. In: Revista Română de Interacțiune Om-Calculator, 3(2), pp. 81-96.

Narayanan, V., Arora, I., Bhatia, A. (2013). Fast and Accurate Sentiment Classification Using An Enhanced Naive Bayes Model. In: Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Li, B., Weise, T., Yao, X. (eds.), Ideal 2013, LNCS 8206, Springer, pp. 194-201.

Ngoc Phuoc, A. V., Popescu, O., Caselli, T. (2014). FBK-TR: SVM for semantic relatedness and corpus patterns for RTE. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval-2014), Dublin, Ireland.

Niculae, V., Zampieri, M., Dinu, L. P., and Ciobanu, A. (2014). Temporal text ranking and automatic dating of texts. In: Proceedings of EACL.

Niculescu, A. (2009). Aspecte ale latinei creștine românești. In: Dacoromania, no. 1, Cluj-Napoca, p. 26.

Nigam, K., Lafferty, J., McCallum, A. (1999). Using Maximum Entropy for Text Classification. In: Proceedings of IJCAI-99. Workshop on Machine Learning for Information Filtering, pp. 61–67.

Nitti, J. and Kasten, L. (1997). The Electronic Texts and Concordances of Medieval Navarro-Aragonese Manuscripts. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

O'Neill, J. (1999). Electronic Texts and Concordances of the Madison Corpus of Early Spanish Manuscripts and Printings. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Oehrle, R., Bach, E. and Wheeler, D. (eds.) (1988). Categorical Grammars and Natural Language Structures, Reidel, Dordrecht.

Ogden, C. K. şi Richards, A. (1923). The Meaning of Meaning; Routledge a. Kegan Paul, Londra.

Onelli, C., Proietti, D., Seidenari, C., Tamburini, F. (2006). The DiaCORIS project: a diachronic corpus of written Italian. In: Proceedings of the 5th International Conference on Language Resources and Evaluation LREC 2006, Genova, pp. 1212-1215.

Palmer, F. R. (1982). Semantics, 2nd edition, Cambridge: Cambridge University Press.

Pană Dindelegan, G., Zafiu, R., Dragomirescu, A., Nicula, I., Nicolae, A., Esher, L. (2015). Diachronic Variation in Romanian, Cambridge Scholars Publishing, UK.

Paradis, C. (2003). Is the notion of linguistic competence relevant in Cognitive Linguistics. Annual Review of Cognitive Linguistics, 1, pp. 247-271.

Paradis, C. (2012). Lexical Semantis. In: The Encyclopedia of Applied Linguistics, ed. Chapelle, C.A. Oxford, UK: WileyBlackwell, pp. 3357-3356.

Perek, F. (2014). Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics ACL-2014 (Short Papers), Baltimore, Maryland, USA, pp. 309-314.

Petcu, M. (2000). Tipologia presei românești, Institutul European, Iași.

Petcu, M. (2009). Istoria presei din România și referențialul european. In: Revista Română de Sociologie, nr. 3-4, București, pp. 277-288.

Petic, M. and Gîfu, D. (2014). Transliteration and Alignment of Parallel Texts from Cyrillic to Latin in *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), N. Calzolari (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.), European Language Resources Association (ELRA), pp. 1819-1823.

Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, MA: MIT Press.

Pollard, C. and Sag, I. A. (1994). Head-driven Phrase Structure Grammar, Chicago: University of Chicago Press.

Pollard, C. and Sag, I. A. (1987). Information-Based Syntax and Semantics. Volume I. CSLI Lecture Notes, number 13. CSLI Publications, Stanford University.

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

Popescu, O. and Strapparava, C. (2014). Time corpora: Epochs, opinions and changes. Knowledge-Based Systems.

Popescu, O. and Strapparava, C. (2015). Semeval-2015 task 7: Diachronic text evaluation. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June 4-5, 2015, Denver, Colorado, pp. 870-878.

Preda, I. (1992). Îmbogățirea lexico-semantică a limbii române actuale (Cu privire la perioada post-decembristă). In: LR, XLI, 1992, nr. 9, pp. 483-490 (I); LR, 1992, nr. 10, pp. 541-549 (II); LR, 1992, nr. 11-12, pp. 589-595 (III); LR, XLII, 1993, nr. 1, pp. 19-28 (IV).

Pustejovsky, J. (1991). The Generative Lexicon. In: Computational Linguistics, 17(4),

pp. 409-441.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, MA: MIT Press.

Pustejovsky, J. and B. Boguraev. (1993). Lexical Knowledge Representation and Natural Language Processing. In: Artificial Intelligence, 63, pp. 193-223.

Pustejovsky, J. and Bergler, S. (eds.). (1992). Lexical Semantics and Knowledge Representation. Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin.

Puşcariu, S. (1913). Dicţionarul limbii române (DA), Întocmit şi publicat după îndemnul şi cu cheltuiala Maiestăţii Sale regelui Carol I, Tomul I, Partea 1, A–B, Bucureşti, Librăriile Socec & Comp. şi C. Sfetea.

Pütz, M. and Sicola, L. (eds.) (2010). Cognitive processing in second language acquisition. Amsterdam & Philadelphia: John Benjamins.

Rama, T and Borin, L. (2014). 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.

Ratnaparkhi, A. (1997). A Linear Observed Time Statistical Parser Based on Maximum Entropy Models. In: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, Brown University, Providence, Rhode Island.

Ratnaparkhi, A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. thesis, UPenn.

Răduică, G. & Răduică, N. (1995). Dictionarul presei românești (1731-1918), Ed. Științifică, București.

Rebedea, T., Chiru, C.-G., Trăușan-Matu, Ș. (2008). Portal web de știri autonom bazat pe prelucrarea limbajului natural. In: Lucrările celei de a 4-a Conferințe Naționale de Interacțiune Om-Calculator – RoCHI 2008, Marhan, A. and Juvina, I. (eds.). Ed. MATRIX ROM, București, pp. 85-92.

Ren, H., Ji, D., Wan, J. (2010). WHU at TAC 2009: A Tri-categorization Approach to Textual Entailment Recognition. In: Proceedings of the Second Text Analysis Conference (TAC 2009) November 16-17, 2009 National Institute of Standards and Technology Gaithersburg, Maryland, USA.

Rennie, J. D. M. , Shih, L., Teevan, J., Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: Proceedings of the 12th International Conference on Machine Learning (ICML-03), AAAI Press.

Resnik, P., Broman Olsen, M. and Diab, M. (1999). The Bible as a Parallel Corpus: Annotating the ‘Book of 2000 Tongues’. Computers and the Humanities 33, pp. 129-153.

Rhaman, K., (2010). Plagiarism Detection Using Semantic Analysis, PhD thesis at BRAC University, Dhaka, Bangladesh.

Rissanen, M. (2008). 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. pp. 53-68.

Rivasseau, J.-N. (2003). Understanding and applying lda model to first-order markov chains. Univ. of british columbia, canada, technical report, Univ. of British Columbia, Canada, Canada.

Robertson, S. E., and Sparck Jones, K. (1976). Relevance weighting of search terms, Journal of the American Society for Information Science, Vol. 27, pp. 129–146.

Robinson, P. and Ellis, N. C. (2008). Conclusion: Cognitive Linguistics, Second Language Acquisition and L2 Instruction - Issues for Research. In Handbook of cognitive linguistics and second language acquisition, edited by P. Robinson and N. C. Ellis. London: Routledge.

Rodriguez, C. (2005). The ABC of Model Selection: AIC, BIC and the New CIC. Bayesian Inference and Maximum Entropy Methods in Science and Engineering.

Roquette, M. L. (2007). Substanţa istorică. In: S. Moscovici and F. Buschini (coords.), Metodologia ştiinţelor socioumane, Editura Polirom, Iași.

Saeed, J. I. (2003). Semantics. Introducing linguistics, 2, Malden, MA: Blackwell Pub.

Sagot, B. (2008). WordNet Libre du Francais (WOLF). INRIA, Paris.

Sagot, B., Darja, F. (2008). Building a free French wordnet from multilingual resources. In: Ontolex 2008, Marrakech, Maroc, p. 6.

Saint-Dizier, P. and Viegas, E. (Eds.). (1995). Computational Lexical Semantics. Cambridge University Press, New York.

Sala, M. (coord.) (1988). Vocabularul reprezentativ al limbilor romanice, București.

Salton, G. (1971). The Smart Retrieval System. Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs.

Salton, G., Yang, C.-S., and Wong, A. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), pp. 613-620.

Sánchez, M. N., Herrera, M. T., and Zabía, M. P. (2003). Textos medievales misceláneos. Hispanic Seminary of Medieval Studies, Ltd. Madison, Wisconsin.

Sánchez-Marco, C., Boleda, G. Fontana, J. M., Domingo, J. (2010). Annotation and representation of a diachronic corpus of Spanish. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 1-6.

Saussure, F. de (1916). Cours de linguistique générale, Paris, Payot (transl. in 1998, Curs de lingvistica generală, Iași, Polirom).

Saussure, Ferdinand de (1916/1959). Course in General Linguistics. New York: The Philosophical Library.

Saussure, F. de (1922). Cours de linguistique générale. 2nd edition. Paris: Payot. Trans. by R. Harris, F. de Saussure, Course in General Linguistics, London: Duckworth, 1983).

Scannel, K. P. (2006). Machine translation for closely related language pairs. In: Proceedings of the Workshop Strategies for developing machine translation for minority languages, pp. 103-109.

Scannel, K. P. (2014). Statistical models for text normalization and machine translation. In Proceedings of the First Celtic Language Technology Workshop, Dublin, Ireland, August 23, pp. 33-40.

Schaff, A. (1966). Introducere în semantică, Editura Ştiinţifică, Bucureşti, pp. 228-330.

Seche, L. and Seche, M. (1997). Dicționarul de sinonime al limbii române, Univers enciclopedic, București.

Seche, M. (1969). Schiţă de istorie a lexicografiei româneşti, vol. II, Bucureşti, Editura Ştiinţifică, pp. 36-39.

Seche, M. and Seche, L. (2002) Dicționar de antonime, Ed. Litera Internațional, 2002.

Simionescu, R. (2016). Inference based and statistical approaches in Natural Language Processing - applications to morphology and syntax, PhD thesis, „Alexandru Ioan Cuza" University of Iași. (*to be published*).

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

Sinclair, J. M. (ed.). (1987). Collins COBUILD English language dictionary (1st ed.). London: Williams Collins Sons & Co Ltd.

Singleton, D. (2000). Language and the lexicon: An Introduction. London: Arnold.

Sperber, H. (1922). Ein Gesetz der Bedeutungsentwicklung. Zeitschrift fuir Deutsches Altertum 59, pp. 49-82.

Stern, G. (1931). Meaning and Change of Meaning. With special reference to the English language. Göteborg: Elanders Boktryckeri Aktiebolag. (Repr. Bloomington & London: Indiana Univ. Press, 1968).

Stoichițoiu, A. (1992). Împrumuturi "necesare" și împrumuturi "de lux" în limbajul publicistic actual. In: Comunicările "Hyperion", Păcurariu, D. (coord.), Ed. Hyperion XXI, Bucuresti, pp. 169-176.

Stoichițoiu-Ichim, A. (1996). Observații privind influența engleză în limbajul publicistic actual. In: LL, XLVI, 1996, II, p. 37-46 (I); LL, XLVI, 1996, III-IV, p. 25-34 (II).

Sukhareva, M. and Chiarcos, C. (2014). 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, pp. 11-20.

Sundermeyer, M., Schluter, R., and Ney, H. (2012) LSTM neural networks for language modeling. In: INTERSPEECH.

Szabo, L.-V. (1999). Libertate si comunicare în lumea presei. Principii, norme, reguli. Limbajul mass-media, Ed. Amarcord, Timișoara.

Szymanski, T. and Lynch, G. UCD: Diachronic Text Classification with Character, Word, and Syntactic N-grams, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), United States, 2015.

Șăineanu, L. (1887). Încercare asupra semasiologiei limbei române. Studii istorice despre transițiunea sensurilor, București.

Șpac, I. (2002). Revista Viaţa Basarabiei 1932-1944. Cercetări bibliografice. Studiu documentar-informativ, Bons offices, Chișinău.

Șpac, I. (2006). Revista Arhivele Basarabiei (1929-1936), Studiu documentar-informativ, Bons offices, Chişinău.

Șuteu, F. (1974). Varianta standard în ierarhia stilistică a limbii, LR, XXIII.

Șuteu, F. (1976). Influenţa ortografiei asupra pronunţării literare româneşti, Editura Academiei Române, Bucureşti.

Tai, K. S., Socher, R., Manning, C. D. (2015). Improved Semantic Representation From Tree-Structured Long Short-Term Memory Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, July 26-31, 2015, Beijing, China, pp. 1556–1566.

Tan, L. and Ordan, N. (2015). USAAR-CHRONOS: Crawling the Web for Temporal Annotations. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, pp. 846-850.

Toma, Săftoiu, E. (1987). SINAL: un sistem de dialog în limbaj natural. In: Buletinul Român de Informatică, Bucureşti.

Tomasello, M. (2003). Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

Trandabăț, D, Irimia, E., Barbu-Mititelu, V., Cristea, D., Tufiș, D. (2012). The Romanian Language in the Digital Age / Limba română în era digitală, in White Paper Series, Eds. Georg Rehm and Hans Uszkoreit, Berlin, Springer.

Trandabăț, D. and Gîfu, D. (eds.) (2016). Linguistic Linked Open Data. Springer.

Tufiş, D. (1982). Sistemul SDLR pentru dialog în limba română. In: Buletinul Român de Informatică, Bucureşti.

Tufiş, D. (1989). It Would Be Much Easier If WENT Were GOED. In: Somers, H., McGee Wood, M. (eds.), Computational Linguistics, Association for Computational Linguistics, Manchester.

Tufiş, D. (1990). Paradigmatic Morphology Learning. In: Computers and Artificial Intelligence, VEDA Publishing House, Bratislava.

Tufiș, D. (1996). Resurse lingvistice computaționale pentru limba română: trecut, prezent și viitor. In: Tufiș, D. (ed.), Limbaj și Tehnologie, Ed. Academiei Române, București, 1996, pp. 59-66.

Tufiş, D. (2008). Ro-WordNet: ontologie lexicală pentru limba română, Academica, XVIII (208-209), pp. 30–34.

Tufiș, D. (2009). Paradigmatic Morphology and Subjectivity Mark-up in the RO-WordNet Lexical Ontology. In: Teodorescu, H.-N., Watada, J, and Jain, L. C. (eds.), Intelligent Systems and Technologies – Methods and Aplications, SCI 217, pp. 161-179, Springer Verlag.

Tufiș, D. and Barbu, E. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets in Proceedings of the 4th LREC Conference, Lisabona, pp. 1067-1068.

Tufiș, D. and Cristea, D. (1985). IURES: A Human Engineering Approach to Natural Language Question Answering. In: Bibel, W., Petfoff, B. (eds.), Artificial Intelligence: Systems, Applications, Methodology, North Holland.

Tufiș, D. and Cristea, D. (2002). Ro-BALKANET – ontologie lexicalizată în context multilingv pentru limba română. In: Limba Română în Societatea Informațională – Societatea Cunoașterii, Ed. Expert, Tufiș, D., Filip, F. Gh. (coord.), pp. 139-166.

Tufiş, D. and Dumitrescu, C. (1990). MORPHO - A Dictionary Management System. In: Proceedings of the fifth International Seminar on Database and Knowledge Systems, Mamaia.

Tufiș, D., Cristea, D., Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal of Information Science and Technology*, Romanian Academy, Bucharest, Romania, Dan Tufis (ed.) Special Issue on

BalkaNet, July, 7(1-2), pp. 9-43.

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

Tufiș, D., Slama-Cazacu, T. (1994). An Overview of Research and Development inLanguage Technology in Romania. In: Proceedings of Conference Awareness Days on Language & Technology, Luxembourg.

Țincu, B. (1974). Transilvania (1868-1900). In: O. Papadima, Reviste literare românești din ultimele decenii ale secolului al XIX-lea, Ed. Academiei Republicii Socialiste România, București.

Țurcanu, I. (2004). Istoricitatea istoriografiei. Observații asupra scrisului istoric basarabean. Chişinău: ARC.

Ullmann, S. (1952/1966). Precis de semantique française. Francke, Berna.

Ullmann, S. (1962). Semantics: An Introduction to the Science of Meaning.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer Verlag, New York.

Vapnik, V. and Cortes, C. (1995). Support vector networks. In: Machine Learning, vol. 20, pp. 273-297.

Vintilă-Rădulescu, I. (2002). Resurse lingvistice pentru limba romană elaborate la Institutul de Lingvistică „Iorgu Iordan". In: *Limba Română în Societatea Informațională – Societatea Cunoașterii*, Tufiș, D., Filip, F. Gh. (coord.), Ed. Expert, București, pp. 21-33.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous time dynamic topic models. In: Proceedings of the International Conference on Machine Learning.

Wang, X. and McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 424-433.

Wang, X., McCallum, A., and Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining.

Wierzbicka, A. (1996). Semantics: primes and universals. Oxford: Oxford University Press.

Wijaya, D. T. and Yeniterzi, R. (2011). Understanding Semantic Change of Words over Centuries. In: Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web, pp. 35-40.

Winkelmolen, F. and Mascardi, V. (2011). Statistical Language Identification of Short

Texts. In: Proceedings of ICAART 2011, vol. 1-Artificial Intelligence, pp. 498-503.

Wittgenstein, L. (1991). Tractatus logico-philosophicus. Transl., introductive note, and notes by Alexandru Surdu, Ed. Humanitas, București.

Wittgenstein, L. (1993). Caietul albastru. Transl. by Lucia Dumitru, Mircea Flonta and Adrian-Paul Iliescu. Introd. note by Mircea Flonta. Bucureşti: Humanitas, pp. 50-51.

Yuan, S. and Qian, Z. (2015). Tibetan-Chinese Cross Language Text Similarity Calculation Based on LDA Topic Model. In: The Open Cybernetics & Systemics Journal, vol. 9, pp. 2911-2919.

Zamfir, C. & Filipescu, I. (coords.) (2015). Sociologie românească 1900-2010. O istorie socială. Ed. Eikon.

Zampieri, M., Ciobanu, A. M., Niculae, V., Dinu, L. P. (2015). AMBRA: A Ranking Approach to Temporal Rext Classification. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June 4-5, 2015, Denver, Colorado, pp. 851-855.

Zazo, R., Lozano-Diez, A., Gonzalez-Dominguez, J., Toledano, D. T., Gonzalez-Rodriguez, J. (2016). Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks. In: PLoS One; 11(1) (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4732772/).

Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009). ANNIS: a search tool for multi-layer annotated corpora. In: Mahlberg, M., González-Díaz, V., Smith, C. (eds.), Proceedings of the Corpus Liguistics Conference, CL2009, University of Liverpool, UK, 20-23 July 2009 (http://edoc.hu-berlin.de/oa/conferences/reS4Xo05sncZc/PDF/29i8VTIzYfT3M.pdf).

Zhu, X. J., Ghahramani, Z, and Lafferty, J. (2003). Semi-supervised Learning using Gaussian Fields and Harmonic Functions. ICML 2003.

The PhD student activity (1st Oct. 2013 – 1st Oct. 2016)

2016

Papers and conferences

- **Gîfu, D.** and Săcălean, L. *Building a Romanian Diachronic Lexicon*, at the 7th International Conference EDU-WORLD 2016, 4-5 Nov. 2014, University of Piteşti, Romania (*accepted*).
- Dascălu, M., **Gîfu**, D., Trăușan-Matu, Ș. *What Makes your Writing Style Unique? Significant Differences between Two Famous Romanian Oratorsi*, at the 8th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI 2016, September 28-30, 2016, Halkidiki, Greece, (impact factor = 0.302) - Class C (*accepted*).
- Cristea, D., Pistol, I., **Gîfu, D.** and Anechitei, D. *Networking Readers: Using Semantic and Geographical Links to Enhance e-Books Reading Experience*, at the 2nd Workshop on Social Media and the Web of Linked Data, RUMOUR 2016, at the 8th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI 2016, September 28-30, 2016, Halkidiki, Greece, (impact factor = 0.302) - Class C (*accepted*).
- **Gîfu, D**. *An Investigation of the Romanian Words Age*, at the 2nd Young Researchers in Artificial Intelligence, TICIA - 2016, 1-3 Aug. 2016, associated to the International Symposium on INnovations in Intelligent SysTems and Applications, INISTA 2016, Sinaia, Romania (*accepted*).
- **Gîfu, D**. *Diachronic Analysis Using a Statistical Model*, at the Mathematical Foundations of Informatics Conference, MFOI 2016, 2nd edition, 25-29 July 2016, Chișinău, Republic of Moldova (*accepted*).
- **Gîfu, D**. *Diachronic Evaluation of Newspapers Language between Different Idioms*, at the first Workshop on Natural Language Meets Journalism, NLPMJ-2016, held at the International Joint Conference on Artificial Intelligence, IJCAI-16, 9-15 July 2016, New York, USA - Class A* (*accepted*).

- **Gîfu, D**. *The Chronology of Old Romanian Words*, at the 3rd International Conference Globalization, Intercultural Dialogue and National Identity, GIDNI 2016, Language and Discourse section, 19-20 May 2016, Târgu-Mureș, Romania (*to be published*).
- **Gîfu, D.** and Cioca, M. *Linguistic Interpretation of Speech Errors*, at the International Conference on Computers Communications and Control, 6th edition, ICCCC2016, 10-14 May 2016, Băile Felix, Oradea, România. In: 2016 6th International Conference on Computers Communications and Control (ICCCC), IEEE Xplrore Library, DOI: 10.1109/ICCCC.2016.7496750, pp. 130-134.
- **Gîfu, D.** and Simionescu, R. *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 – Class B (*to be published*).
- Colhon, M, Cristea, D. and **Gîfu, D.** *Discovering Semantic Relations within Nominals*. In: Trandabăț D. and Gîfu, D. (eds.), Linguistic Linked Open Data. 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop, Sibiu, Romania, July 13-25, 2015, Revised Selected Papers, © Springer-Verlag Berlin Heidelberg, 2016, pp. 85-100.
- Cristea, D., **Gîfu, D.**, Niculiță, M., Pistol, I., and Sfirnaciuc, D. *A Mixed Approach in Recognising Geographical Entities in Texts*. In: Trandabăț D. and Gîfu, D. (eds.), Linguistic Linked Open Data. 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop, Sibiu, Romania, July 13-25, 2015, Revised Selected Papers, © Springer-Verlag Berlin Heidelberg, 2016, pp. 49-64.
- **Gîfu, D.**, *The Analysis of Diachronic Variation in Romanian Print Press*. In: Proceedings of the First PhD Symposium on Sustainable Ultrascale Computing Systems, NESSUS PhD 2016, Carretero, J, Garcia Blas, J., Petcu, D. (eds.), Computer Architecture, Communications, and Systems Group (ARCOS), Spain, Feb. 2016, pp. 49-53, ISBN: 978-84-608-6309-0.
- **Gîfu, D.** and Cioca, M. *Innovative Methods for Business Education Using Isotope Linking on Anonymous Readers' Comments*, at the Balkan Region Conference on Engineering and Business Education. Volume 1, Issue 1, ISSN (Online) 2391-8160, DOI: 10.1515/cplbu-2015-0023, January 2016.

Books

- Trandabăț, D. and **Gîfu, D.** (eds.). Linguistic Linked Open Data, 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop, Sibiu, Romania, July 13-25, 2015, Revised Selected Papers, © Springer-Verlag Berlin Heidelberg, 2016, ISSN 1865-0929, ISSN

1865-0937 (electronic), ISBN 978-3-319-32941-3, ISBN 978-3-319-32942-0 (eBook), DOI 10.1007/978-3-319-32942-0.

Member in Organizing Committees

The 4th BringITon! 2016 Workshop, 18-19 Nov. 2014, Iaşi, Romania.

The 11th International Conference Linguistic Resources and Tools for processing of the Romanian language, ConsILR-2015, 27-29 Oct. 2016, Iași, Romania.

The 2nd Workshop on Social Media and the Web of Linked Data, RUMOUR-2016, held at the 8the International Conference on Computational Collective Intelligence, ICCCI2016, 28-30 September 2016, Halkidiki, Greece.

The 14th International Conference on Human-Computer Interaction, RoCHI 2016, 8-9 Sept. 2016, Iași, Romania.

The European Language Resource Coordination Workshop, ELRC, hosted by European Commission Representation in Romania, 23 Mar. 2016, Bucharest, Romania.

Member in Program Committees

The 9th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2017, 3-5 April 2017, Kanazawa, Japan.

The 4th BringITon! 2016 Workshop, 18-19 Nov. 2014, Iaşi, Romania.

The 11th International Conference Linguistic Resources and Tools for processing of the Romanian language, ConsILR-2015, 27-29 Oct. 2016, Iași, Romania.

The 6th International Workshop on Applications of Software Agents WASA-2016, held at the 8the International Conference on Computational Collective Intelligence, ICCCI 2016, 28-30 September 2016, Halkidiki, Greece.

The 2nd RUMOUR-2016 Workshop on Social Media and the Web of Linked Data, held at the 8the International Conference on Computational Collective Intelligence, ICCCI 2016, 28-30 September 2016, Halkidiki, Greece.

The first Workshop on Natural Language Meets Journalism NLPMJ-2016, held at the International Joint Conference on Artificial Intelligence, IJCAI-16, 9-15 July 2016, New York, USA.

The 2nd Mathematical Foundations of Informatics Conference MFOI-2016, 25-29 July 2016, Chișinău, Moldova.

The 10th International Conference on Language Resources and Evaluation, LREC-2016, 23-28 May 2016, Portorož, Slovenia.

Peer-review

Since 2016 - reviewer of International Journal of Computers & Technology, USA & India, ISSN: 2277-3061 (Online), impact factor 1.213.

Affiliations

Since 2016 - member of the Romanian Group of Interest in Human-Computer Interaction, RoCHI, formally affiliated at ACM SIGCHI (ACM Special Interest Group for Human Computer Interaction) Romania since 7 June 2001.

2015

Papers and conferences

- **Gîfu, D.** *Semantic Particularities of Journalistic Text*, at the 3rd International Conference Literature, Discourse and Multicultural Dialogue, section: Communication, Public Relations, Journalism, LDMD-2015, 3-4 Dec. 2015, "Petru Maior" University of Târgu-Mureș. In: Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue, I. Boldea (coord.), vol. 3, Arhipelag XXI Press, 2015, pp. 182-196, ISBN: 978-606-8624-21-1.
- **Gîfu, D.**, Pistol, I., Cristea, D. *Annotation Conventions for Geographical Relations*, at the 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania. In: Proceedings of The 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiș (eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, 2015, pp. 67-78, ISSN 1843-911X, indexed by CNCSIS.
- Colhon, M., **Gîfu, D.** and Cristea, D. *The QuoVadis Storytelling*, at the 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania. In: Proceedings of The 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiș (eds.), "Alexandru

Ioan Cuza" University Publishing House, Iaşi, 2015, pp. 93-108, ISSN 1843-911X, indexed by CNCSIS.

- Bejan, I.-M., Iftene, A. and **Gîfu, D.** *A Lexical Discourse Analysis Framework*, at the 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania. In: Proceedings of The 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiș (eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2015, pp. 111-124, ISSN 1843-911X, indexed by CNCSIS.
- Dascălu, M. and **Gîfu, D.** *Evaluating the Complexity of Online Romanian Press*, at the 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania. In: Proceedings of The 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiș (eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2015, pp. 149-162, ISSN 1843-911X, indexed by CNCSIS.
- **Gîfu, D.** and Cioca, M. *Innovative Methods for Business Education Using Isotope Linking on Anonymous Readers' Comments*, at the 3rd International Engineering and Technology Education Conference, IETEC'15, and 7th Balkan Region Conference on Engineering and Business Education, BRCEBE, 1-4 Nov. 2015, Sibiu, Romania.
- **Gîfu, D.** *Particularities of Romanian Journalistic Language in the 19th century*, at the International Interdisciplinary Doctoral Conference, IIDC 2015, first edition, 25-26 Sept. 2015, Bucharest, Romania (*to be published*).
- **Gîfu, D.** *Romanian Language - the Analysis of Diachronic Variation in Novels*, at the Young Researchers in Artificial Intelligence, TICIA, 5-8 Sept. 2015, Cluj-Napoca, Romania.
- **Gîfu, D.** *Contrastive Diachronic Study on Romanian Language*, at the European Workshop on Foundation of Informatics, FOI 2015, first edition, 24-29 Aug. 2015, Chișinău, Moldova, in Proceedings FOI-2015, S. Cojocaru, C.Gaindric (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, 2015, pp. 296-310, ISBN 978-9975-4237-3-1, indexed by Mathematical Reviews and Zentralblatt MATH.
- Tufiș, D., Cristea, D. and **Gîfu, D.** The EUROLAN-2015 Summer School „Linguistic Linked Open Data" (LLOD)". In: Academica, nr. 8-9, Aug-Sept. 2015, Publishing House of the Romanian Academy, Bucharest, Romania, pp. 36-38, ISSN 1220-5737

- **Gîfu, D.** and Iliescu, A. *Analysis of Bridge Anaphora Across Novel*, at the 6th International Conference EDU-WORLD 2014, 7-9 Nov. 2014, University of Piteşti, Romania. In: Procedia – Social and Behavioral Sciences, vol. 180, 5 May 2015, E. Soare and C. Langa (eds.), pp. 1474-1480, ISSN: 1877-0428, indexed by Elsevier.
- **Gîfu, D.** *Use of Isotopes in the Online Commentators' Categorization*, at the 2nd International Conference Globalization, Intercultural Dialogue and National Identity, GIDNI 2015, Section: Communication, Public Relations, Journalism, 28-29 May 2015, "Petru Maior" University of Târgu-Mureș. In: Proceedings of the International Conference Globalization, Intercultural Dialogue and National Identity, I. Boldea (coord.), vol. 2, Arhipelag XXI, Târgu-Mureș, pp. 189-200, ISBN 978-606-93692-5-8.
- Sălăvăstru, A. and **Gîfu, D.** *Annotating Geographical Entities*, at the 16th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2015, 14-20 Apr. 2015, Cairo, Egypt - Class B.
- **Gîfu, D.** *Creating of Parallel Lexicons for Romanian and Moldovan*, at the Communication, Context, Interdisciplinarity Congress, "Petru Maior" University, Târgu-Mureş, 23-24 Oct. 2014. In: Communication, Context, Interdisciplinarity, I. Boldea (coord.), vol. 3, Ed. "Petru Maior" University, Târgu-Mureş, 2015, section: Language and discourse, pp. 617-626, ISSN 2069-3389.
- **Gîfu, D.** *Pragmatics of Discursive Politeness*. In: Globalization and intercultural dialogue: multidisciplinary perspectives, Communication and PR section, I. Boldea (coord.), Ed. Arhipelag XXI, Târgu-Mureș, Romania, vol. 1, 2014, pp. 77-87, ISBN 978-606-93691-3-5 - indexed by ISI Web of Knowledge/Thomson Reuters.
- **Gîfu, D.** and Topor, R. *Recognition of Discursive Verbal Politeness*, at the 11th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2014, Venice, Italy, 27-29 Oct. 2014, published in Natural Language Processing and Cognitive. Proceedings 2014, B. Sharp and R. Delmonte (eds.), by the Publisher De Gruyter, Italy, 2015, pp. 123-134, ISBN: 978-1-5015-1042-7 - Class C.
- **Gîfu, D.** *Can We Talk About Humor in the Religious Speech?*, at the International Symposium Integrare europeană/ identitate națională; plurilingvism/ multiculturalitate – limba și cultura română: evaluări, perspective, Romanian Academy – Iași branch, 2013. Published in: European Integration/National Identity. Pluringualism/Multiculturality. The Romanian language and culture: evaluations, perspectives. Proceedings, Luminița Botoșineanu, Ofelia Ichim (eds.), Ed. Aracne, Rome, Italy, „Danubiana" Collection, Dec. 2014, pp. 91-100, ISBN 978-88-548-7812-9, indexed by CEEOL, EBSCO.

Books

- **Gîfu, D.**, Trandabăț, D., Cristea, D., Tufiș, D. (eds.). Proceedings of The 11th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania, "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2015, 215p., ISSN 1843-911X - indexed by CNCSIS.

Books chapters

- Cristea, D., **Gîfu, D.**, Colhon, M., Diac, P., Bibiri, A., Mărănduc, C., and Scutelnicu, L.-A. Chapter - Quo Vadis: A Corpus of Entities and Relations. In: Language, Production, Cognition, and the Lexicon. Text, Speech and Language Technology, Part VI - Language Resouces and Langauge Engineering, Nuria Gala, Reinhard Rapp and Gemma Bel-Enguix (eds.), vol. 48/2015, New York, USA, pp. 505-543 - indexed in Springer, ISI Web of Science/Thomson Reuters.

Mobilities

20-22 Oct. 2015 - Innovative, Connect, Transform - ICT 2015: Digital Agenda for Europe, Lisbon, Portugal.

01 Sept. 2014-30 June 2015 - Erasmus Mundus Action 2 EMERGE Project (2011 – 2576 / 001 – 001 - EMA2) – an exchange mobility at the "Alecu Russo" Bălți State University, Republic of Moldova. Field of study: Computer Science, Computational Linguistics, Semantic and Pragmatic aspects of the discourse. The indicative budget is: 460.000.000 EUR. Grade: PhD student.

Member in Organizing Committees

The 11th International Conference Linguistic Resources and Tools for processing of the Romanian language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania.

The 12th EUROLAN-2015 Summer School on Linguistic Linked Open Data, 13 - 25 July 2015, Sibiu, Romania.

The RUMOUR-2015 Workshop on Social Media and the Web of Linked Data, 18 July 2015, Sibiu, Romania - a satellite event of EUROLAN-2015 Summer School on Linguistic Linked Open Data.

Member in Program Committees

The 9th IEEE International Conference on Open Source Systems and Technologies, ICOSST-2015, and associated Open Source Series Workshop (OSSW-2015), 17-19 Dec. 2015, Lahore, Pakistan.

The 11th International Conference Linguistic Resources and Tools for processing of the Romanian language, ConsILR-2015, 26-27 Nov. 2015, Iași, Romania.

The RUMOUR-2015 Workshop on Social Media and the Web of Linked Data, 18 July 2015, Sibiu, Romania - a satellite event of EUROLAN-2015 Summer School on Linguistic Linked Open Data.

Editorial board

Since 2015 - member of the Society of Electronics and Computer Engineering Journal, Hyderabad, Telangana, India.

Peer-review

Since 2015 - reviewer of Computer Science and Information Technology, Horizon Research Publishing, USA, ISSN: 2331-6063 (Print) and ISSN: 2331-6071 (Online).

Affiliations

Since 2015 - member of the Romanian Association for Artificial Intelligence, ARIA, București, Romania.

Keynote speaker

Social Media in Our Lives. Online Profile Features, at the 9th IEEE International Conference on Open Source Systems and Technologies and associated Open Source Series Workshop (OSSW-2015), 17-19 Dec. 2015, Lahore, Pakistan.

2014

Papers and conferences

- **Gîfu, D.** & Cioca, M. *Detecting Emotions in Comments on Forums*. In: International Journal of Computers Communications and Control, I. Dzitac, F.G. Filip, M.-J. Manolescu (eds.), vol. 9, no. 6/2014 (Dec.), Agora University Editing House, pp. 694-702, ISSN 1841-9836, eISSN 1841-9844 - indexed by ISI Web of Science/Thomson Reuters (impact factor = 0.694).

- **Gîfu, D.** and Vasilache, G. *A Language Independent Named Entity Recognition System*. In: Proceedings of The 10th International Conference Linguistic Resources and Tools for Processing The Romanian Language, ConsILR-2014, 18-19 Sept. 2014, Craiova, Romania, Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiş (eds.), "Alexandru Ioan Cuza" University Publishing House, Iaşi, 2014, pp. 181-188, indexed by CNCSIS.
- **Gîfu, D.** *The Emotional Orientation*. In: Proceedings of the third Conference of Mathematical Society of the Republic of Moldova, "IMCS-50", 19-23 August 2014, Chișinău, Republic of Moldova, pp. 511-516, ISBN 978-9975-68-244-2, indexed by Mathematical Reviews and Zentralblatt MATH.
- Petic, M. and **Gîfu, D.** *Transliteration and Alignment of Parallel Texts from Cyrillic to Latin*. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), N. Calzolari (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.), European Language Resources Association (ELRA), 26-31 May 2014, Reykjavik (Iceland), ISBN 978-2-9517408-8-4 – indexed by DBLP and ISI Web of Science/Thomson Reuters - Class B.
- **Gîfu, D.**, Scutelnicu, L.-A., Cristea, D. *Humour and Non-humour in Religious Discourse*. In: Proceedings of the 2nd Edition Workshop on Language Resources and Evaluation for Religious Texts, 31 May 2014, Reykjavik (Iceland). A post-conference workshop co-hosted at LREC'2014 (Language Resources and Evaluation Conference, Claire Brierly, Majdi Sawalha, Eric Atwell (eds.), pp. 30-35 - indexed by ISI Web of Science/Thomson Reuters - Class C.
- **Gîfu, D.** *Pragmatics of Discursive Politeness*. In: Proceedings of the International Conference Globalization, Intercultural Dialogue and National Identity, Iulian Boldea (ed.), Volume no. 1, 2014, Section: Communication and Public Relations, 29-30 May 2014 at the "Petru Maior" University of Târgu-Mureș, Arhipelag XXI, Târgu-Mureș, pp. 77-87, ISBN 978-606-93691-3-5 - indexed by ISI Web of Science/Thomson Reuters.
- **Gîfu, D.** & Cioca, M. *Sentiment Analysis in Social Media*. In: Abstract of ICCCC Papers, 4(2014), I. Dziţac, F.G. Filip, M.-J. Manolescu (eds.), Agora University & Information Science and Technology Section of the Romanian Academy: 5th International Conference on Computers, Communications & Control, ICCCC2014, Romania, Oradea & Băile Felix, May 6-10, 2014, ISSN 1844-4334 - (impact factor = 0.746).
- **Gîfu, D.** and Scutelnicu, L.-A. *Natural Language Complexity. Sentiment Analysis*, at the 1st edition, 5-6 Dec. 2013, "Petru Maior" University in collaboration with The ALPHA

Institute for Multicultural Studies, Târgu-Mureș. In: Studies On Literature, Discourse and Multicultural Dialogue, Iulian Boldea (ed.), Ed. ARHIPELAG XXI, Târgu-Mureș, 2014, pp. 944-954, ISBN 978-606-93590-3-7 - indexed by ISI Web of Science/Thomson Reuters.

Books

- Alboaie, L., Cristea, D., Forăscu, C., **Gîfu, D.**, Iftene, A. (eds.). BringITon! 2014 Catalogue, "Alexandru Ioan Cuza" University Publishing House, Iași, 2014, 68p., ISSN 2285-0929, ISSN-L: 2285-0929.
- Smarandache, F., Vlăduțescu, Ș, **Gîfu, D.**, Țenescu, A. (eds.), Topical communication uncertainties, Sitech, Craiova and & Zip Publishing, Ohio/SUA, 2014, 230p., ISBN 978-606-11-4225-5 & 9781599739830.

Mobilities

18-20 Mar. 2014 - IDEALIST-H2020 “Proposal Check Event” (calls H2020-ICT-2014-1), Cologne, Germany.

Member in Organizing Committees

The 3rd BringITOn 2014 Workshop, 14-15 Nov. 2014, Iaşi, Romania.

The 10th International Conference Linguistic Resources and Tools for processing of the Romanian language, ConsILR-2014, 18-19 Sept. 2014, Craiova, Romania.

Member in Program Committees

The 8th IEEE International Conference on Open Source Systems and Technologies, ICOSST-2014, and associated Open Source Series Workshop (OSSW-2014), 18-20 Dec. 2014, Lahore, Pakistan.

The 3rd BringIT On 2014 Workshop, 14-15 Nov. 2014, Iaşi, Romania.

The 10th International Conference Linguistic Resources and Tools for Processing of the Romanian Language, ConsILR-2014, 18-19 Sept. 2014, Craiova, Romania.

The 9th International Conference on Language Resources and Evaluation LREC-2014, 26-31 May 2014, Reykjavik, Iceland.

Editorial board

Since 2014 - member of the The Global Electrical Engineers Journal, Avanti Publishers, Karachi, Pakistan & Windsor, Connecticut, USA, ISSN (online): 2410-0412.

Since 2014 - member of the Social Sciences and Educational Research Review, Sitech Publishing House – Craiova, ISSN: 2392-9683, ISSN-L: 2392-9683.

Since 2014 - member of the Computer Engineering and Information Technology Journal CEITJ, Indore, Madhya Pradesh, India, ISSN: 2394-8434, ZDB-ID: 2814237-8.

Keynote speaker

Sentiment Analysis Impact in Business, at the 8th IEEE International Conference on Open Source Systems and Technologies and associated Open Source Series Workshop (OSSW-2014), 18-20 Dec. 2014, Lahore, Pakistan.

Print Press from Republic of Moldova - a diachronic approach (Presa scrisă din Republica Moldova - o abordare diacronică), at the 4th Scientific Colloquium Current Orientations in the Doctoral Research (Orientări Actuale în Cercetarea Doctorală), 18 Dec. 2014, "Alecu Russo" Bălți State University, Republic of Moldova.

2013

Papers and conferences

- Delmonte, R., **Gîfu, D.** and Tripodi, R. *Extracting Opinion and Factivity from Italian political discourse*. In: Proceedings of NLPCS 2013 10th International Workshop on Natural Language Processing and Cognitive Science, Bernadette Sharp and Michael Zock (Eds.), Marseille, France, 15-16 Oct. 2013, pp. 162-176 - Class C.
- **Gîfu, D.** and Cristea, D. *Monitoring and predicting journalistic profiles*, at the 5th International Conference on Computational Collective Intelligence Technologies and Applications, ICCCI 2013; Craiova; Romania; 11-13 Sept. 2013 - Class C.

Published in: Computational collective intelligence: technologies and applications (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 8083 LNAI, 2013, C. Bădică, N.T. Nguyen and M. Brezovan (eds.), © Springer-Verlag Berlin Heidelberg, 2013, ISSN 0302-9743, ISBN 978-3-642-40494-8 - indexed by SCOPUS, DBLP and ISI Web of Science (impact factor = 0.302).

Books

- **Gîfu, D.** Temeliile Turnului Babel. O perspectivă integratoare asupra discursului politic, Ed. Academiei Române, Bucureşti, 2013, 284p., ISBN 978-973-272-1827.

Mobilities

06-08 Nov. 2013 - Information and Communication Technologies - ICT 2013: Create, Connect, Grow, Vilnius, Lithuania.

Member in Organizing Committees

The 10th Scientific School in Natural Language Processing, EUROLAN-2013, 11-19 October 2013, Ipotești, Romania.