An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian

Alina Maria Ciobanu, Liviu P. Dinu
Faculty of Mathematics and Computer Science, University of Bucharest
Center for Computational Linguistics, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

In this paper we propose a computational method for determining the orthographic similarity between Romanian and related languages. We account for etymons and cognates and we investigate not only the number of related words, but also their forms, quantifying orthographic similarities. The method we propose is adaptable to any language, as far as resources are available.

1 Introduction

Language relatedness and language change across space and time are two of the main questions of the historical and comparative linguistics (Rama and Borin, 2014). Many comparative methods have been used to establish relationships between languages, to determine language families and to reconstruct their proto-languages (Durie and Ross, 1996). If grouping of languages in linguistic families is generally accepted, the relationships between languages belonging to the same family are periodically investigated. In spite of the fact that linguistic literature abounds in claims of classification of natural languages, the degrees of similarity between languages are far from being certain. In many situations, the similarity of natural languages is a fairly vague notion, both linguists and non-linguists having intuitions about which languages are more similar to which others. McMahon and McMahon (2003) and Rama and Borin (2014) note that the computational historical linguistics did not receive much attention until the beginning of the 1990s, and argue for the necessity of development of quantitative and computational methods in this field.

1.1 Related Work

According to Campbell (2003), the methods based on comparisons of cognate lists and sound correspondences are the most popular approaches employed for establishing relationships between languages. Barbançon et al. (2013) emphasize the variety of computational methods used in this field, and state that the differences in datasets and approaches cause difficulties in the evaluation of the results regarding the reconstruction of the phylogenetic tree of languages. Linguistic phylogeny reconstruction proves especially useful in historical and comparative linguistics, as it enables the analysis of language evolution. Ringe et al. (2002) propose a computational method for evolutionary tree reconstruction based on a "perfect phylogeny" algorithm; using a Bayesian phylogeographic approach, Alekseyenko et al. (2012), continuing the work of Atkinson et al. (2005), model the expansion of the Indo-European language family and find support for the hypothesis which places its homeland in Anatolia; Atkinson and Gray (2006) analyze language divergence dates and argue for the usage of computational phylogenetic methods in the question of Indo-European age and origins. Using modified versions of Swadesh's lists[1], Dyen et al. (1992) investigate the classification of Indo-European languages by applying a lexicostatistical method.

The similarity of languages is interesting not only for historical and comparative linguistics, but for machine translation and language acquisition as well. Scannell (2006) and Hajič et al. (2000) argue for the possibility of obtaining a better translation quality using simple methods for very closely related languages. Koppel and Ordan (2011) study the impact of the distance between languages on the translation product and conclude that it is directly correlated with the ability to distinguish translations from a given source language from non-translated text. Some genetically related languages are so similar to each other, that

[1] http://www.wordgumbo.com/ie/cmp/iedata.txt

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1047–1058,
October 25-29, 2014, Doha, Qatar.

speakers of such languages are able to communicate without prior instruction (Gooskens, 2007). Gooskens et al. (2008) analyze several phonetic and lexical predictors and their conclusion is that lexical similarity can be seen as a predictor of language intelligibility. The impact of language similarities in the process of second language acquisition is argued by the contrastive analysis hypothesis, which claims that where similarities between the first and the second language occur, the acquisition would be easier compared with the situation in which there were differences between the two languages (Benati and VanPatten, 2011).

1.2 Our Approach

Although there are multiple aspects that are relevant in the study of language relatedness, such as the orthographic, phonetic, syntactic and semantic differences, in this paper we focus only on the orthographic similarity. The orthographic approach relies on the idea that sound changes leave traces in the orthography, and alphabetic character correspondences represent, to a fairly large extent, sound correspondences (Delmestri and Cristianini, 2010).

In this paper we propose an orthographic similarity method focused on etymons (direct sources of the words in a foreign language) and cognates (words in different languages having the same etymology and a common ancestor). In a broadly accepted sense, the higher the similarity degree between two languages, the closer they are.

One of our motivations is that when people encounter a language for the first time in written form, it is most likely that they can distinguish and individualize words which resemble words from their native language. These words are probably either inherited from their mother tongue (etymons), or have a common ancestor with the words in their language (cognates).

Our first goal is, given a corpus C, to automatically detect etymons and cognates. In Section 2 we propose a dictionary-based approach to automatically extract related words, and a method for computing the orthographic similarity of natural languages. Most of the traditional approaches in this field focus either on etymology detection or on cognate identification, most of them reporting results only on small sets of cognate pairs (usually manually determined lists of about 200 cognates, for which the cognate judgments are made by human experts (Rama and Borin, 2014)). Our approach implies a detailed investigation which accounts not only for the number of related words, as it is usually done in lexicostatistics (where the relationships between languages are determined based on the percentage of related words), but also for their forms, quantifying orthographic similarities. We employ three string similarity metrics for a finer-grained analysis, as related words in different languages do not have identical forms and their partial similarity implies different degrees of recognition and comprehensibility. For example, the Romanian word *lună* (*moon*) is closer to its Latin etymon *luna* than the word *bătrân* (*old*) to its etymon *veteranus*, and the Romanian word *vânt* (*wind*) is closer to its French cognate pair *vent* than the word *castel* (*castle*) to its cognate pair *château*.

In this paper we investigate the orthographic similarity between Romanian and related languages. Romanian is a Romance language, belonging to the Italic branch of the Indo-European language family, and is of particular interest regarding its geographic setting. It is surrounded by Slavic languages and its relationship with the big Romance kernel was difficult. Besides general typological comparisons that can be made between any two or more languages, Romanian can be studied based on comparisons of genetic and geographical nature, participating in numerous areally-based similarities that define the Balkan convergence area. Joseph (1999) states that, regarding the genetic relationships, Romanian can be studied in the context of those languages most closely related to it and that the well-studied Romance languages enable comparisons that might not be possible otherwise, within less well-documented families of languages. The position of Romanian within the Romance family is controversial (McMahon and McMahon, 2003): either marginal or more integrated within the group, depending on the versions of the cognate lists that are used in the analysis.

In Section 3.1 we apply our method on Romanian in different stages of its evolution, running our experiments on high-volume corpora from three historical periods: the period approximately between 1642 and 1743, the second half of 19th century (1870 - 1889), and the present period. In Section 3.2 we make use of a fourth corpus, Europarl, with a double goal: on the one hand, to check if degrees of similarity between Romanian

and other languages in the present period are consistent across two different corpora, and on the other hand, to investigate whether there are differences between the overall degrees of similarity obtained for the entire corpus and those obtained in various experiments at sentence level. The conclusions of our paper are outlined in Section 4.

2 Methodology and Algorithm

In this section we introduce a technique for determining the orthographic similarity of languages. In order to obtain accurate results, we investigate both etymons and cognates. First, we automatically identify etymons and cognates, then we measure the distances between related words, and finally we compute the overall degrees of similarity between pairs of languages. We also applied this method for investigating the mutual intelligibility of the Romance languages, and preliminary results are presented in (Ciobanu and Dinu, 2014b).

2.1 Similarity Method

Let $C = \{w_1, w_2, ..., w_{N_{words}}\}$ be a corpus in L_1 and let L_2 be a related language. We assume, without any loss of generality, that the elements of C are ordered such that $C_L = \{w_1, w_2, ..., w_{N_{lingua}}\}$ is the subset of C containing all the words that have an etymon or a cognate pair in L_2. We use the following notations: N_{words} is the number of token words in C, N_{lingua} is the number of token words in C_L, λ is the empty string and x_i is the etymon or cognate pair of w_i in L_2. Given a string distance Δ, we define the distance between L_1 and L_2 (non-metric distance), with frequency support from corpus C, as follows:

$$\Delta(L_1, L_2) = 1 - \frac{N_{lingua}}{N_{words}} + \frac{\sum_{i=1}^{N_{lingua}} \Delta(w_i, x_i)}{N_{words}} \quad (1)$$

Hence, the similarity between languages L_1 and L_2 is defined as follows:

$$Sim(L_1, L_2) = 1 - \Delta(L_1, L_2) \quad (2)$$

2.2 Algorithm

We present here the algorithm based on linguistic relationships detection and string similarity methods for determining the orthographic similarity between languages, with frequency support from corpora in the source language. This algorithm, represented in Figure 2, is applicable to any language. After a preprocessing phase, which is detailed in Subsection 2.2.1, we analyze words and begin by identifying their etymologies.

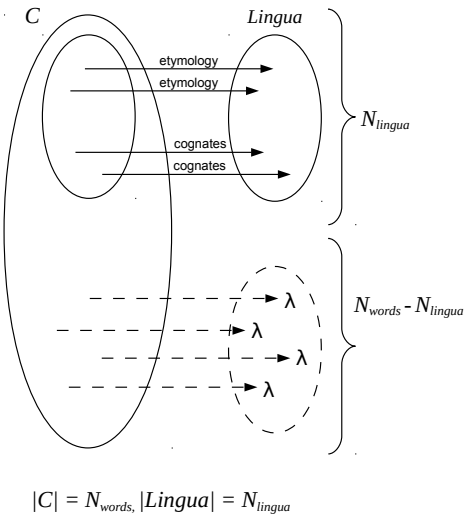

Figure 1: Schema for determining the orthographic similarity between related languages with frequency support from corpus *C*.

Corpus	#words		#stop words		#lemmas
	token	type	token	type	type
Parliament	22,469,290	162,399	14,451,178	214	40,065
Eminescu	870,828	65,742	565,396	212	21,456
Chronicles	253,786	28,936	170,582	193	8,189
RVR	2,464	2,464	124	124	2,252

Table 1: Statistics for the Romanian datasets.

2.2.1 Preprocessing

Given a corpus C, we start by preprocessing the text.

Step 1. Data Cleaning. We perform basic word segmentation, using whitespace and punctuation marks as delimiters and we lower-case all words. We remove from the datasets tokens that are irrelevant for our investigation, such as dates, numbers and non-textual annotations marked by non-alphanumeric characters.

Step 2. Stop Words Removal. We focus on analyzing word content and, in order to obtain relevant results, we remove stop words from the datasets. We use the lists of stop words for Romanian provided by the Apache Lucene[2] text search engine library. In Table 3.1 we list the total number of stop words from each corpus.

[2]http://lucene.apache.org

Step 3. Lemmatization. We use lemmas for identifying words' definitions in dictionaries and for computing adequate distances between words and their cognates or etymons. We use the Dexonline[3] machine-readable dictionary to lemmatize Romanian words.

Step 4. Diacritics Removal. Many words have undergone transformations by the augmentation of language-specific diacritics when entering a new language. From an orthographic perspective, the resemblance of words is higher between words without diacritics than between words with diacritics. For example, the orthographic distance is higher for the Romanian word *amiciţie* (*friendship*) and its French cognate pair *amitié* than for their corresponding forms without diacritics, *amicitie* and *amitie*. For this reason, in this step of our procedure we create two versions of each dataset, with and without diacritics, in order to further investigate the influence of the diacritics on the cross-language orthographic similarity. In Romanian, 5 diacritics are used today: *ă, î, â, ş, ţ*.

2.2.2 Relationships Identification

Step 1. Etymology Detection. For most words, etymological dictionaries offer a unique etymology, but when more options are possible for explaining a word's etymology (there are words whose etymology was and remains difficult to ascertain), dictionaries may provide multiple alternatives. For example, the Romanian word *parlament* (*parliament*) has a double etymology: French (with the etymon *parlement*) and Italian (with the etymon *parlamento*). We account for all the given etymological hypotheses, enabling our method to provide more accurate results.

For determining words' etymologies we use the Dexonline machine-readable dictionary, which is an aggregation of over 30 Romanian dictionaries. By parsing its definitions, we are able to automatically extract information regarding words' etymologies and etymons. The most frequently used pattern is shown below.

```
<abbr class="abbrev"
    title="limba language_name">
    language_abbreviation </abbr>
<b> origin_word </b>
```

As an example, we provide below an excerpt from a Dexonline entry which uses this pattern to specify the etymology of the Romanian word *capitol* (*chapter*), which has double etymology: Latin (with the etymon *capitulum*) and Italian (with the etymon *capitolo*).

```
<b> CAPÍTOL </b>
<abbr class="abbrev"
    title="limba italiana"> it. </abbr>
<b> capitolo </b>
<abbr class="abbrev"
    title="limba latina"> lat. </abbr>
<b> capitulum </b>
```

Step 2. Cognate Identification. Cognates are words in different languages having the same etymology and a common ancestor. The methods for cognate detection proposed so far are mostly based on orthographic/phonetic and semantic similarities (Kondrak, 2001; Frunza et al., 2005), but the term "cognates" is often used with a somewhat different meaning, denoting words with high orthographic/phonetic and cross-lingual meaning similarity, the condition of common etymology being left aside. We focus on etymology and we introduce an automatic strategy for detecting pairs of

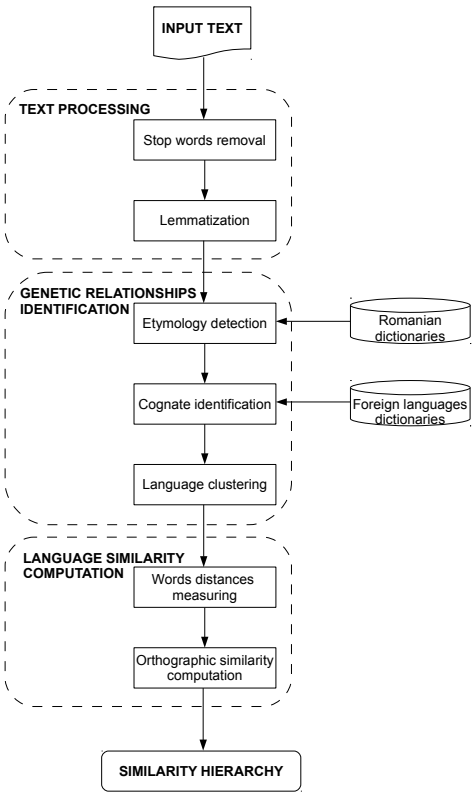

Figure 2: Algorithm for determining the orthographic similarity between related languages.

[3] http://dexonline.ro

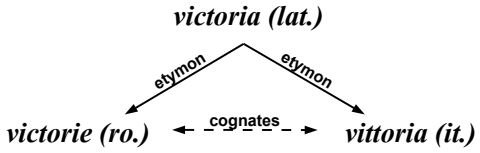

Figure 3: Word-etymon and cognate pairs.

cognates between two given languages, which enables the identification of all cognate pairs for the studied corpus.

Considering a set of words in a given language *L*, to identify the cognate pairs between *L* and a related language *L'*, we first determine the etymologies of the given words. Then we translate in *L'* all words without *L'* etymology. We consider cognate candidates pairs formed of input words and their translations. Using electronic dictionaries, we extract etymology-related information for the translated words. To identify cognates, we compare, for each pair of candidates, the etymologies and the etymons. If they match, we identify the words as being cognates.

In our previous work (Ciobanu and Dinu, 2014a) we applied this method on a Romanian dictionary, while here we extract cognates from Romanian corpora. We identify cognate pairs between Romanian and six other languages: Italian, French, Spanish, Portuguese, Turkish and English. We use electronic dictionaries[4] to extract etymology-related information and Google Translate[5] to translate Romanian words. We are restricted in our investigation by the available resources, but we plan to extend our method to other related languages as well. We selected these six languages for the following reason: the first four in our list are Romance languages, and our intuition is that there are numerous words in these languages which share a common ancestor with Romanian words. We investigate the cognate pairs for Turkish because many French words were imported in both Romanian and Turkish in the 19th century, and we believe that accounting for Romanian-Turkish cognates would provide a more accurate result for the similarity of these languages. As for English, we decided to investigate the cognate pairs for this language in order to analyze to what extent the influence of English on Romanian increases across time. In Table 2 we report the number of Romanian words having an etymon or a cognate pair in the six related languages.

Step 3. Evaluation. In order to evaluate our automatic method for extracting etymology-related information and for detecting related words, we excerpt a sample of 500 words for each of the considered languages (Romanian, French, Italian, Spanish, Portuguese, Turkish and English). The samples are drawn using a proportionate stratification sampling method with regard to the length of the lemmas in our datasets. We manually determine the etymologies of the words in the samples, and we compare these results with the automatically obtained etymologies. We compute the accuracy for etymology extraction for each language, and we obtain the following results: 95.8% accuracy for Romanian, 97.8% for Italian, 96.8% for French, 96.6% for Spanish, 97.0% for Portuguese, 96.0% for Turkish and, finally, 97.2% for English.

Language	Relationship	Corpus			
		Parliament	Eminescu	Chronicles	RVR
French	cognates	192,275	13,074	3,139	43
	etymons	15,665,865	484,668	89,946	1,203
Italian	cognates	1,660,588	40,491	2,743	100
	etymons	9,234,710	348,948	77,633	957
Spanish	cognates	4,616,528	119,627	9,942	355
	etymons	4,411,707	212,106	65,336	482
Portuguese	cognates	4,378,354	115,309	15,755	324
	etymons	3,477,285	156,908	55,991	435
Turkish	cognates	1,401,569	33,070	2,332	113
	etymons	331,863	24,115	11,985	69
English	cognates	4,347,302	146,377	21,966	296
	etymons	625,596	17,328	6,799	56

Table 2: Number of Romanian token words having etymons or cognate pairs in related languages.

2.2.3 Linguistic Distances

Various approaches have been previously employed for assessing the orthographic distance or similarity between related words. Their performance has been investigated and compared (Frunza et al., 2005; Rama and Borin, 2014), but a clear conclusion cannot be drawn with respect to which method is the most appropriate for a given task. We employ three metrics to determine the orthographic similarity between related words. In Subsection 3.1.2 we investigate to what extent the similarity scores computed with each of these metrics differ, and whether the differences are statistically significant. We use the following metrics:

[4] Italian: http://www.sapere.it/sapere/dizionari
French: http://www.cnrtl.fr
Spanish: http://lema.rae.es/drae
Portuguese: http://www.infopedia.pt/lingua-portuguesa
Turkish: http://www.nisanyansozluk.com
English: http://www.collinsdictionary.com

[5] http://translate.google.com

- **LCSR:** The longest common subsequence ratio (Melamed, 1995) is the longest common subsequence of two strings u and v divided by the length of the longer word. We subtract this value from 1, in order to obtain the distance between two words.

- **EDIT:** The edit distance (Levenshtein, 1965) counts the minimum number of operations (insertion, deletion and substitution) required to transform one string into another. We use a normalized version of the edit distance, dividing it by the length of the longest string.

- **RD:** The rank distance (Dinu and Dinu, 2005) is used to measure the similarity between two ranked lists. A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, ..., n$. Let S be a set of ranking results, $\sigma \in S$. $\sigma(i)$ represents the rank of object i in the ranking result σ. The rank distance is computed as: $RD(\sigma, \tau) = \sum_{i=1}^{n} |\sigma(i) - \tau(i)|$. The ranks of the elements are assigned from bottom up, i.e. from n to 1, using the Borda count method (de Borda, 1781). The elements which do not occur in any of the rankings receive the rank 0. To extend the rank distance to strings, we index each occurrence of a given letter a with a_k, where k is the number of its previous occurrences, and then apply the rank distance on the new indexed strings, which become rankings in this situation. In order to normalize it, we divide the rank distance by the maximum possible distance between two strings u and v (Dinu and Sgarro, 2006): $\Delta_{max}(u, v) = |u|(|u|+1)/2 + |v|(|v|+1)/2$.

3 Experiments and Results

In this section we present the results obtained by applying our method for determining orthographic similarity on Romanian datasets.

To our knowledge, only basic lexicostatistical methods (generally based on different dictionaries or versions of the representative vocabulary of Romanian) which compute the percentage of words with a given etymology have been applied for determining the relationships between Romanian and related languages. Because of the difficulty of setting the bounds between the basic lexicon and the remaining words, Graur (1968) uses in his experiments three concentric versions of the basic Romanian lexicon. Dinu (1996) reevaluates the etymology detection for the three versions of the basic Romanian lexicon and reclassifies the lexical material. He argues against grouping together all the words with Slavic origins, without differentiation between Old Slavic and languages such as Bulgarian, Russian, Ukrainian and Polish. Sala (1988) builds a version of the representative vocabulary of Romanian comprising 2588 words, which we use in our experiments as well.

3.1 Romanian Evolution

We apply our similarity method on high-volume Romanian corpora from three distinct historical periods of time, with different cultural, economical, political and social contexts. In Table we report statistics for these corpora and for the basic Romanian lexicon.

3.1.1 Data

The first corpus consists of the transcription of the parliamentary debates held in the Romanian Parliament from 1996 to 2007 (Grozea, 2012). The second corpus consists of the publishing works of Mihai Eminescu (Eminescu, 1980-1985), the leading Romanian poet. His works provide an insightful description of the period between 1870 and 1889, with respect to its cultural, economical, social and political aspects, including some major events in the Romanian history. Many researchers consider that Eminescu had a crucial influence on Romanian, his contribution to modern language development being highly appreciated. The third corpus dates back to the period approximately between 1642 and 1743, the beginning period of the Romanian writing. Miron Costin, Grigore Ureche and Ion Neculce are Romanian chroniclers whose main works follow one another in creating one of the most detailed and valuable descriptions of Moldavia in that period, "Letopiseţul Ţării Moldovei". Along with them, Dimitrie Cantermir contributed to the early development of the Romanian writing, having written what is considered to be the first attempt at a socio-political novel ("Istoria Ieroglifică", 1703-1705). Their chronicles account for social, cultural, economical and political events with the purpose of recreating historical periods of time. We also use the basic Romanian lexicon (Sala, 1988), abbreviated RVR, for our experiments. The Dexonline machine-readable dictionary, which we use for determining the etymologies for the Romanian words, aggregates defini-

Language	Parliament					Eminescu					Chronicles					RVR				
	%words	D		ND		%words	D		ND		%words	D		ND		%words	D		ND	
French	70.6	45.5	46.0	48.3	48.8	57.2	35.2	36.1	37.2	38.2	36.7	20.3	21.1	22.3	23.1	50.6	30.3	31.4	32.2	33.3
Latin	63.7	40.2		42.0		59.9	34.6		36.6		44.9	24.2		25.7		56.5	34.0		37.3	
Italian	48.5	28.1	33.4	29.1	34.5	44.7	26.9	30.2	27.9	31.2	31.7	19.6	20.3	20.7	21.4	41.4	23.4	26.2	25.2	28.0
Spanish	40.2	9.2	24.9	10.7	27.0	38.1	10.9	21.2	12.9	23.7	29.7	11.9	15.1	13.9	17.2	32.5	9.0	19.5	9.9	21.0
Portuguese	35.0	8.3	22.1	9.5	24.0	31.3	9.6	18.5	11.3	21.0	28.3	12.2	16.3	13.9	18.4	29.3	8.6	17.4	9.4	18.9
English	22.1	2.2	14.0	2.2	14.2	18.8	1.1	9.9	1.2	10.1	11.3	1.3	5.9	1.3	6.2	14.3	1.6	10.3	1.6	10.4
Provencal	17.7	9.6		9.8		20.7	11.3		11.6		21.8	13.0		13.4		16.8	9.7		10.5	
German	9.2	5.8		5.9		6.9	4.5		4.6		4.9	2.4		2.4		10.2	6.3		6.6	
Turkish	7.7	0.9	5.4	0.9	5.6	6.6	1.7	4.5	1.7	4.7	5.6	2.9	3.7	3.1	3.9	7.4	1.6	5.0	1.8	5.3
Russian	5.9	3.7		4.0		6.5	4.0		4.4		7.5	4.3		4.9		9.0	5.4		6.2	
Catalan	5.9	3.3		3.4		9.0	4.8		5.1		11.2	5.9		6.4		8.4	4.6		4.9	
Greek	4.8	2.9		3.0		6.0	3.6		3.7		4.5	2.6		2.7		4.6	2.5		2.6	
Albanian	4.8	2.6		3.0		6.7	3.7		4.0		9.1	4.9		5.3		8.4	4.2		4.8	
Bulgarian	4.0	2.6		3.0		7.4	4.7		5.5		10.6	6.8		7.8		11.8	7.2		8.4	
Slavic	4.9	2.3		2.5		6.6	3.4		3.8		12.1	6.5		7.7		9.8	5.0		5.7	
Old Slavic	3.8	2.2		2.7		6.1	3.3		4.3		11.9	6.8		8.7		9.5	5.2		6.0	
Hungarian	2.9	1.8		2.0		5.1	2.9		3.3		7.5	4.3		4.7		7.4	3.7		4.6	
Ruthenian	2.4	1.6		2.0		4.7	3.0		3.7		6.0	3.7		4.4		4.5	2.4		3.0	
Serbian	2.6	1.4		1.6		5.8	3.0		3.4		8.9	5.0		5.5		8.6	5.2		6.0	
Sardinian	1.7	1.0		1.0		3.3	1.7		1.8		4.0	2.0		2.1		2.6	1.4		1.5	

Table 3: Results for the Romanian datasets. In the *D* and *ND* columns we provide the average degrees of similarity for the datasets with and without diacritics. For languages for which we determine cognate pairs (besides etymons), we report both versions of the results, before and after cognate identification. In the *%words* column we provide the percentage of words having an etymon or a cognate pair in each language. The results are ordered according to the ranking of similarity for the corpus comprising the parliamentary debates after identifying cognates and with diacritics included.

tions from over 30 dictionaries ranging from 1927 to the present time and contains archaisms and obsolete words (which are marked accordingly); therefore, we are able to identify etymologies for words in all used corpora.

3.1.2 Results

In this subsection we present and analyze the main results drawn from our research. In Table 3 we list the output of our method for each corpus, with and without diacritics[6]. We report the similarity between Romanian and related languages, providing the average value of the three metrics used. In the *%words* column we provide, for each corpus, the percentage of words having an etymon or a cognate pair in a given language (the typical measure used in lexicostatistical comparison, i.e., the 0 distance function). The results for the Romanian datasets are plotted in Figure 4 and Figure 5.

Cognate Influence. Table 3 and Figure 4 present the gain obtained by cognate analysis. Accounting for cognates leads to an increase of similarity between Romanian and Spanish and Portuguese with almost 300%, and between Romanian and Italian with almost 20%. Another spectacular increase of closeness is for Turkish, which draws closer to Romanian with more than 500% by using the cognate gain. The degree of similarity is not given by the contribution of words inherited in Romanian from Turkish (about 1%), but by the pairs of shared cognates. Both Romanian and Turkish borrowed words from French massively towards the end of the 19th century. Thus, most pairs of Romanian-Turkish cognates have common French ancestors, and words in Romanian and Turkish which resemble are actually loans from the same French words. We also notice a significant increase in similarity between Romanian and English in the modern period. This increase is natural and probably arises for the similarity between English and most of the other languages as well. We notice that this increase is due preponderantly to the cognate pairs. Most of the Romanian-English cognates have a Romance common ancestor (78.4% Latin, 4.2% French, 3.4% Italian), and 11.8% have a Greek common ancestor, counted at lemma level on the corpus comprising the parliamentary debates.

Romanian Evolution. Some significant results can be observed in the evolution of Romanian: the degrees of similarity between Romanian and all the Romance languages has increased significantly from the Chronicles period until today. Besides them, German is the only language to which Romanian drew closer (a possible explanation might be the fact that, after the establishment of Germans

[6]The complete ranking of similarity is available online at http://nlp.unibuc.ro/resources/rosim.pdf.

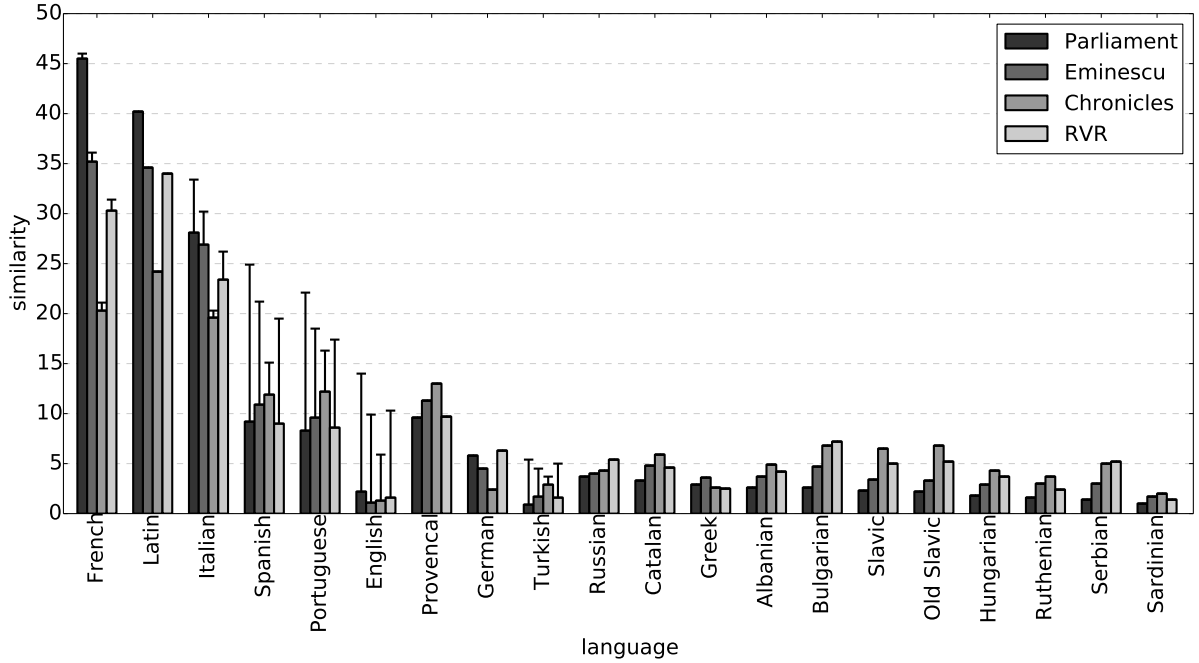

Figure 4: Degrees of similarity for the Romanian datasets. For French, Italian, Spanish, Portuguese, Turkish and English, the values obtained after the cognate identification phase are also plotted.

in Banat and Transylvania, many German words entered the basic Romanian lexicon). On the contrary, the similarity between Romanian and almost all the Slavic languages decreased in the same period. Russian is the only Slavic language which preserved its degree of similarity with Romanian (being, in the 2000s, the only Slavic language among the top 10 most closely related languages with Romanian, on the ninth position, with a degree of similarity of less than 4%). In the 18th and 19th centuries, the transition to the Latin alphabet and the desire to restore Romanian's Latin appearance contributed to the decrease of the Slavic influence (Gheţie, 1978). In fact, all Slavic influences in the 2000s sum up to 8.9%, in contrast with Latin influences, reaching 61.8%. Greek is the only language which reaches its peak regarding the similarity with Romanian in the 19th century (due to the brief Phanariot dominance in the 19th century). Therefore, Romanian preserved its Latin character all along, and the influence of the non-Latin languages on Romanian (overestimated in some works) was in fact not so significant. This fact supports Darwin's theory (Darwin, 1859), which states that the genealogy of languages is consistent with the genealogy of the nations (analyzed based on DNA similarity).

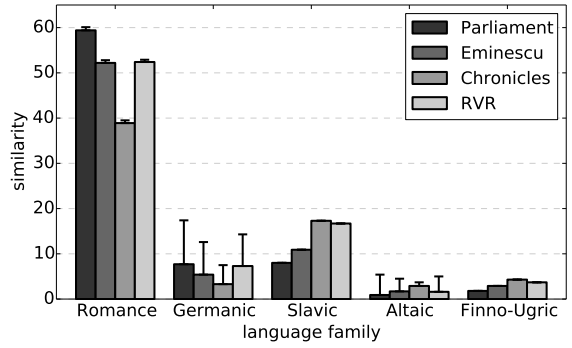

Figure 5: Degrees of similarity for the language families. Iranian and Baltic have a degree of similarity of less than 0.5.

Orthographic Metrics. In order to compare the similarity scores computed with the three metrics used, we conduct hypothesis tests (Sheskin, 2003) to determine whether the differences between the results obtained with each metric are statistically significant. We extract a sample of 5,000 words and we compute the pairwise differences between the similarity scores. Using the R v3.1.0 software environment for statistical computing (R Core Team, 2014), we perform the one-way ANOVA F-test, with the null hypothesis $\mathcal{H}_0$: $\mu_{EDIT} = \mu_{LCSR} = \mu_{RD}$ (where μ_Δ is the mean of the val-

ues computed with the Δ metric) and the alternative hypothesis $\mathcal{H}_a$: *not all* $\mu_{EDIT}, \mu_{LCSR}, \mu_{RD}$ *are equal*. Since the p-value of 2.88e-05 is much smaller than the 0.05 significance level, we have very strong evidence to reject the null hypothesis that the mean computed values for the three metrics are all equal. Further, we perform post-hoc comparisons applying pairwise t-tests with Bonferonni correction for the p-value, in order to analyze the differences between the metrics. For each pair of metrics, $p \ll 0.05$. The differences are statistically significant, but we notice that they are small. Applying a two-sampled t-test, we obtain a [0.012, 0.015] confidence interval for the mean difference between EDIT and LCSR, [0.015, 0.018] for EDIT and RD, and [0.001, 0.003] for RD and LCSR, at 95% confidence level. Moreover, computing Spearman's rank correlation coefficient for the rankings obtained by each metric for each dataset, we observe a very high correlation between them ($\rho > 0.98$ for each pair of variables). Thus, we conclude that reporting the average of the three metrics is relevant for our experiments, as differences are small and do not influence the ranking.

3.2 Europarl Experiments

We continue our investigation regarding the similarity of natural languages with two additional experiments. First, we want to see if degrees of similarity between Romanian and other languages in the present period are consistent across two different corpora. In the second experiment we are interested to see if there are differences between the overall degrees of similarity obtained for the entire corpus (the bag-of-words model) and those obtained in various experiments at sentence level. Our main corpus is Europarl (Koehn, 2005). More specifically, we use the portions larger than 2KB collected between 2007 and 2011 from the Romanian subcorpus of Europarl. The corpus is tokenized and sentence-aligned in 21 languages. For preprocessing this corpus, we discard all the transcribers' descriptions of the parliamentary sessions (such as "The President interrupted the speaker" or "The session was suspended at 19:30 and resumed at 21:00").

Exp. #1. In a first step, we apply the methodology described in Section 2 on the entire Europarl corpus for Romanian, using a bag-of-words model for the entire corpus, in which we account for the overall frequencies of the words. In this experiment, as in the previous ones, we cannot detect outliers, i.e., sentences which are unbalanced regarding the etymologies of the comprised words. For this reason, we conduct a second experiment which addresses this potential issue.

Exp. #2. We determine sentence-level orthographic similarity and we aggregate the results to compute the average values for the related languages. In this second experiment, we apply the methodology described in Section 2 for each sentence in the Europarl corpus for Romanian. For each sentence we obtain a ranking of related languages and, in order to obtain a ranking of similarity for the entire corpus, we compute the average degrees of similarity: for each related language, we sum up the degrees of similarity for all the sentences and divide this value by the number of sentences in the corpus.

Exp. #3. Because the interpretation of statistics derived from datasets that include outliers may be misleading, we compute the standard quartiles *Q1, Q2* and *Q3* (Sheskin, 2003) with regard to the length of the sentences. We use the interquartile range $IQR = Q3 - Q1$ to find outliers in the data. We consider outliers the observations that fall below the lower fence $LF = Q1 - 1.5(IQR)$ or above the upper fence $UF = Q3 + 1.5(IQR)$. We apply our methodology again only for the sentences having the length in the $[LF, UF]$ range.

Language	Exp. #1	Exp. #2	Exp. #3	Exp. #4
French	53.1	52.1	52.1	52.8
Latin	44.1	43.6	43.6	44.0
Italian	40.6	39.9	39.9	40.2
Portuguese	33.6	32.9	32.8	33.2
Spanish	27.6	27.3	27.3	26.8
English	16.0	15.7	15.7	15.1
Provencal	10.0	10.1	10.1	9.3
Turkish	6.3	6.2	6.1	5.7
German	5.9	5.8	5.8	5.3
Greek	4.4	4.3	4.3	3.8
Russian	4.2	4.1	4.1	3.6
Catalan	4.1	4.2	4.2	3.5
Old Slavic	3.1	3.2	3.2	2.7
Albanian	3.0	3.1	3.1	2.5
Bulgarian	2.9	2.9	2.9	2.2
Slavic	2.6	2.5	2.5	1.9
Hungarian	2.5	2.4	2.4	1.7
Ruthenian	2.1	2.1	2.1	1.3
Serbian	1.6	1.5	1.5	0.7
Sardinian	1.2	1.2	1.3	0.1

Table 4: Results for Europarl on the entire corpus (Exp. #1) and at sentence level (Exp. #2 - #4).

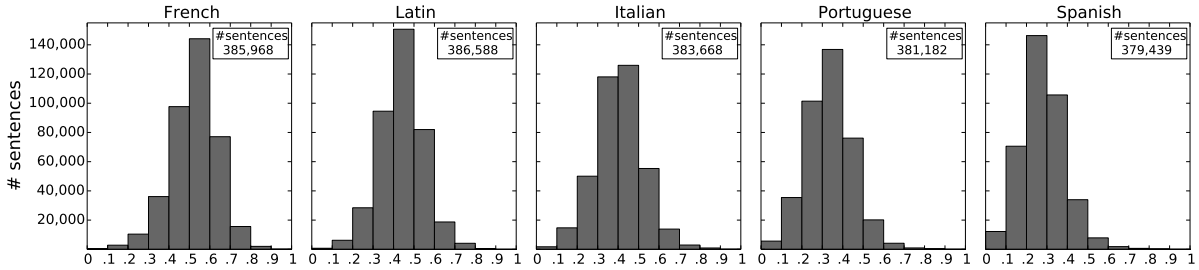

Figure 6: Distribution of the Romanian sentences in Europarl based on their similarity with the top 5 ranked languages. The OX axis represents the degree of similarity normalized to [0,1].

Exp. #4. As a last experiment, for each language L we consider as observations the degrees of similarity between Romanian and L, we discard outliers and we compute the average value of the observations inside the $[LF, UF]$ range. For each language, we determine the distribution of the sentences according to their similarity with the given language (the histograms for the top 5 languages are presented in Figure 6).

In Figure 7 we report the top 20 languages in the ranking of similarity for Europarl, emphasizing the gain obtained by identifying cognates. In Table 4 we report the similarity scores for the top 20 languages in the rankings of similarity for all the 4 experiments: overall similarity for the entire Europarl corpus (Exp. #1), sentence-level similarity (Exp. #2), similarity for the sentences having the length in the $[LF, UF]$ range (Exp. #3), and similarity for the sentences having the similarity between Romanian and each related language in the $[LF, UF]$ range (Exp. #4).

Some remarks are immediate. We observe that the differences between the values obtained for the entire corpus (Exp. #1) and those obtained in various experiments at sentence level (Exp. #2 - #4) are very small (an exception is Exp. #4, for languages whose degrees of similarity with Romanian are of less than 10%). We test the bag-of-words model on two corpora from the same period (the Parliament corpus and Europarl – in Exp. #1) and we notice that the results are consistent across different corpora (0.98 Spearman's ρ). The only significant difference is for Portuguese, which is closer to Romanian as measured on Europarl than on the Parliament corpus.

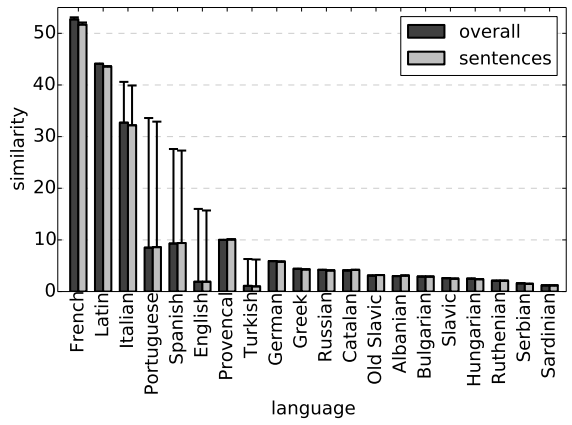

Figure 7: Degrees of similarity for Europarl.

4 Conclusions and Future Work

In this paper we proposed a computational method for determining the cross-language orthographic similarity, with application on Romanian. We investigated etymons and cognates and we conducted a fine-grained analysis of the orthographic similarity between Romanian and related languages. Our results provide a new insight into the classification and evolution of Romanian. We plan to apply our similarity method on a corpus of spoken language, and to extend our analysis to other languages as well, as we gain access to available resources. We further intend to combine our orthographic approach with syntactic and semantic evidence for a wider perspective on language similarity.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. We thank Raluca Vasilache for the help with evaluating the automatic method for detecting related words. The contribution of the authors to this paper is equal. Research supported by CNCS UEFISCDI, project number PNII-ID-PCE-2011-3-0959.

References

Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337(6097):957–960.

Quentin D. Atkinson and Russell D. Gray. 2006. How Old is the Indo-European Language Family? Illumination or More Moths to the Flame? In Peter Forster and Colin Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, chapter 8, pages 91–109. McDonald Institute for Archaeological Research.

Quentin D. Atkinson, Russell D. Gray, Geoff K. Nicholls, and David J. Welch. 2005. From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103(2):193–219.

Francois Barbançon, Steven N. Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods. *Diachronica*, 30(2):143 – 170.

Alessandro G. Benati and Bill VanPatten. 2011. Key Terms in Second Language Acquisition. *International Journal of Applied Linguistics*, 21(2):270–273.

Lyle Campbell. 2003. How to Show Languages are Related: Methods for Distant Genetic Relationship. In Brian D. Joseph and Richard W. Janda, editors, *The Handbook of Historical Linguistics*. Blackwell.

Alina Maria Ciobanu and Liviu P. Dinu. 2014a. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043.

Alina Maria Ciobanu and Liviu P. Dinu. 2014b. On the Romance Languages Mutual Intelligibility. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3313–3318.

Charles Robert Darwin. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray.

Jean-Charles de Borda. 1781. *Mémoire sur les Élections au Scrutin*. Histoire de l'Académie Royale des Sciences.

Antonella Delmestri and Nello Cristianini. 2010. String Similarity Measures and PAM-like Matrices for Cognate Identification. *Bucharest Working Papers in Linguistics*, 12(2):71–82.

Anca Dinu and Liviu P. Dinu. 2005. On the Syllabic Similarities of Romance Languages. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2005*, pages 785–788.

Liviu P. Dinu and Andrea Sgarro. 2006. A Low-Complexity Distance for DNA Strings. *Fundam. Inform.*, 73(3):361–372.

Mihai Dinu. 1996. *Personalitatea Limbii Române*. Cartea Românească.

Mark Durie and Malcolm Ross. 1996. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean Classification: a Lexicostatistical Experiment. *Transactions of the Americal Philosophical Society*, 82(5):1–132.

Mihai Eminescu. 1980-1985. *Opere. Vol IX-XIII. Publicistică*. Editura Academiei Române.

Oana Frunza, Diana Inkpen, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2005*, pages 251–257.

Ion Gheţie. 1978. *Istoria Limbii Române Literare. Privire Sintetică*. Editura Ştiinţifică şi Enciclopedică.

Charlotte Gooskens, Wilbert Heeringa, and Karin Beijering. 2008. Phonetic and Lexical Predictors of Intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2):63–81.

Charlotte Gooskens. 2007. The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. *Journal of Multilingual and Multicultural Development*, 28(6):445.

Alexandru Graur. 1968. *Tendinţele Actuale ale Limbii Române*. Editura Ştiinţifică.

Cristian Grozea. 2012. Experiments and Results with Diacritics Restoration in Romanian. In *Proceedings of the 15th International Conference on Text, Speech and Dialogue, TSD 2012*, pages 199–206.

Jan Hajič, Jan Hric, and Vladislav Kubon. 2000. Machine Translation of Very Close Languages. In *Proceedings of the 6th Conference on Applied Natural Language Processing, ANLC 2000*, pages 7–12.

Brian D. Joseph. 1999. Romanian and the Balkans: Some Comparative Perspectives. In Sheila Embleton, John E. Joseph, and Hans-Joseph Niederehe, editors, *The Emergence of the Modern Language Sciences*. John Benjamins Publishing Company.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit, AAMT 2005*, pages 79–86.

Grzegorz Kondrak. 2001. Identifying Cognates by Phonetic and Semantic Similarity. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL 2001*, pages 1–8.

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT 2011*, pages 1318–1326.

Vladimir I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710.

April McMahon and Robert McMahon. 2003. Finding Families: Quantitative Methods in Language Classification. *Transactions of the Philological Society*, 101(1):7–55.

Dan Melamed. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 184–198.

R Core Team, 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Taraka Rama and Lars Borin. 2014. Comparative Evaluation of String Similarity Measures for Automatic Language Classification. In George K. Mikros and Ján Macutek, editors, *Sequences in Language and Text*. De Gruyter Mouton.

Don Ringe, Ann Taylor, and Tandy Warnow. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Marius Sala. 1988. *Vocabularul Reprezentativ al Limbilor Romanice*. Editura Academiei.

Kevin Scannell. 2006. Machine Translation for Closely Related Language Pairs. In *Proceedings of the Workshop on Strategies for Developing Machine Translation for Minority Languages*, pages 103–107.

David J Sheskin. 2003. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC Press.