Automatic Identification and Production of Related Words for Historical Linguistics

Alina Maria Ciobanu
University of Bucharest
Department of Computer Science
HLT Research Center
alina.ciobanu@my.fmi.unibuc.ro

Liviu P. Dinu
University of Bucharest
Department of Computer Science
HLT Research Center
ldinu@fmi.unibuc.ro

Language change across space and time is one of the main concerns in historical linguistics. In this article, we develop tools to assist researchers and domain experts in the study of language evolution.

First, we introduce a method to automatically determine whether two words are cognates. We propose an algorithm for extracting cognates from electronic dictionaries that contain etymological information. Having built a data set of related words, we further develop machine learning methods based on orthographic alignment for identifying cognates. We use aligned subsequences as features for classification algorithms in order to infer rules for linguistic changes undergone by words when entering new languages and to discriminate between cognates and non-cognates.

Second, we extend the method to a finer-grained level, to identify the type of relationship between words. Discriminating between cognates and borrowings provides a deeper insight into the history of a language and allows a better characterization of language relatedness. We show that orthographic features have discriminative power and we analyze the underlying linguistic factors that prove relevant in the classification task. To our knowledge, this is the first attempt of this kind.

Third, we develop a machine learning method for automatically producing related words. We focus on reconstructing proto-words, but we also address two related sub-problems, producing modern word forms and producing cognates. The task of reconstructing proto-words consists of recreating the words in an ancient language from its modern daughter languages. Having modern word forms in multiple Romance languages, we infer the form of their common Latin ancestors. Our approach relies on the regularities that occurred when words entered the modern languages. We leverage information from several modern languages, building an ensemble

Submission received: 15 July 2018; revised version received: 24 July 2019; accepted for publication: 17 September 2019.

https://doi.org/10.1162/COLI_a_00361

system for reconstructing proto-words. We apply our method to multiple data sets, showing that our approach improves on previous results, also having the advantage of requiring less input data, which is essential in historical linguistics, where resources are generally scarce.

1. Introduction

Natural languages are living ecosystems—they are constantly in contact and, by consequence, they change continuously. Two of the fundamental questions in historical linguistics are the following: *How are languages related?* and *How do languages change across space and time?* (Rama and Borin 2014). Traditionally, both problems have been investigated with comparative linguistics instruments (Campbell 1998). The main idea of the comparative method is to perform a property-based comparison of multiple sister languages in order to infer properties of their common ancestor. For a long period, the comparative reconstruction has been a time-consuming manual process that required a large amount of intensive work. Addressing the first question implies developing methods for identifying cognates. Addressing the second question implies investigating borrowings and analyzing how words evolve from one language into another.

Cognates are words in different languages having the same meaning and a common ancestor. Investigating pairs of cognates is very useful not only in historical and comparative linguistics (in the study of language relatedness [Ng et al. 2010], phylogenetic inference [Atkinson et al. 2005], and in identifying how and to what extent languages changed over time or influenced each other), but also in other research areas, such as language acquisition, bilingual word recognition (Dijkstra, Grootjen, and Schepens 2012), corpus linguistics (Simard, Foster, and Isabelle 1992), cross-lingual information retrieval (Buckley et al. 1997), and machine translation (Kondrak, Marcu, and Knight 2003).

According to Hall (1960), there is no such thing as a "pure language"—a language "without any borrowing from a foreign language." The process by which words enter one language from another is called **linguistic borrowing**. A borrowed word, also called **loanword**, is defined as a "lexical item (a word) which has been 'borrowed' from another language, a word which originally was not part of the vocabulary of the recipient language but was adopted from some other language and made part of the borrowing language's vocabulary" (Campbell 1998).

The unprecedented contact between languages in today's context of high mobility and the explosion of communication tools led to an inherent enrichment of languages by borrowings.[1] *Why* and *how* the borrowing process takes place are fundamental questions that, by their nature, invite experimental perspective (Chitoran 2011). To answer the first question, Campbell (1998, page 59) notes that "Languages borrow words from other languages primarily because of need and prestige." Further, the author states that the result of the borrowing process depends on numerous factors, such as the length and intensity of the contact and the extent to which the populations in question are bilingual. In problems of language classification, distinguishing cognates from borrowings is essential. Although admittedly regarded as relevant factors in the history of a language (McMahon et al. 2005), borrowings bias the genetic classification of the languages, characterizing them as being closer than they actually are (Minett and Wang 2003). According to Gray and Atkinson (2003), correctly determining cognates and borrowings

1 A dictionary of recent words in Romanian (Dimitrescu 1997) counts 4,853 new words entered after 1965, most of them entering the language after 1990.

is essential in the process of phylogenetic inference, as false cognates and unrecognized borrowings could incorrectly increase the degree of similarity between languages. False cognates are more harmful than missing valid cognates in language comparison, because they can lead to incorrect conclusions regarding the genetic relationships between languages (List, Greenhill, and Gray 2017). Thus, the need for discriminating between cognates and borrowings emerges. Heggarty (2012) acknowledges the necessity and difficulty of the task, emphasizing the role of the "computerized approaches."

Reconstructing proto-words, which is central to the study of language evolution, consists of recreating the words in an ancient language from the words in its modern daughter languages. Bouchard-Côté et al. (2013) emphasize the important role this task plays in historical linguistics, because it enables evaluating proposals regarding the phenomenon of language change. Although the main hypothesis in this research problem is that there are regularities and patterns in how words evolved from the ancestor language to its modern daughter languages, there are also words that diverged significantly from their ancestor. Take, for example, the Latin word *umbilicu(lu)s* (meaning *umbilicus*): it evolved into *buric* (Romanian), *nombril* (French), and *umbigo* (Portuguese), three forms that are dissimilar to one another and to the Latin word. Reconstructing proto-words is a challenging task, and several studies have gone beyond the comparative method to automate the process of proto-language reconstruction (Oakes 2000; Bouchard-Côté et al. 2013; Atkinson 2013).

Other closely related research problems are the production of cognates (determining the form of a given word's cognate pair) and modern words (determining the form in which a proto-word evolves in a modern daughter language). We emphasize two research directions that rely on these tasks of producing word forms: diachronic linguistics, which is concerned with language evolution over time, and the study of foreign language learning, which focuses on the learning process and on the influence of the learner's mother tongue in the process of second language acquisition. Producing cognates can also contribute to the task of lexicon generation for poorly documented languages with scarce resources.

This article is organized as follows. In Section 2 we provide an overview of our research. In Section 3 we present previous work on identifying and producing related words. In Section 4 we describe the process of building a data set of related languages. In Section 5 we introduce our methods for identifying cognates and for discriminating between cognates and borrowings. In Section 6 we describe our system for producing related words, with its three sub-tasks: reconstructing proto-words, producing modern word forms, and producing cognates. Finally, in Section 7 we draw conclusions and present several directions for future work.

2. Our Approach

In this article, we propose a series of tools and resources to provide support in computational historical linguistics.[2]

Our first goal is to automatically identify the relationship between words, focusing on cognates and borrowings. More specifically, we propose methods for identifying cognates and for discriminating between cognates and borrowings.

2 An updated Web page with the resources and tools for historical linguistics proposed in this article will be maintained at `http://nlp.unibuc.ro/resources.html`.

First, we develop, implement, and evaluate a dictionary-based approach to identifying cognates based on the etymology of the words. The proposed method has the advantage of creating large databases of cognate pairs, and it can be easily generalized to languages for which electronic dictionaries with etymological information are available. As a case study, we apply this method on Romanian, identify the etymologies of the Romanian words, and determine cognate pairs between Romanian and related languages (such as Italian, French, Spanish, and Portuguese). Using these pieces of information, we build a data set of multilingual cognates for Romanian, and develop a parallel list of over 3,000 cognate sets across five Romance languages with common Latin etymology.

Second, we introduce a method to automatically determine whether two words are cognates. We use an orthographic alignment method that proved relevant for sequence alignment in computational biology (Needleman and Wunsch 1970), but has become popular in natural language processing as well. We use aligned subsequences as features for machine learning algorithms in order to infer rules for linguistic changes undergone by words when entering new languages and to discriminate between cognates and non-cognates. We apply our method on a subset of the automatically extracted data set of cognates presented in Section 4 for Romanian and four related Romance languages: Italian, French, Spanish, and Portuguese.

Third, we investigate the task of discriminating between cognates and borrowings. The challenge and importance of this task is emphasized by Heggarty (2012, page 122) as follows:

> What solution is there, then, if we can neither ignore the problem of distinguishing cognates from loanwords, nor overcome it in the many cases where we do not have the necessarily linguistic knowledge to do so? There is in fact a possibility: to sidestep the question entirely at the data analysis stage, and simply to identify which forms are judged to be somehow 'correlate' with each other—whether by specialists in those languages, or more objectively by computerized approaches.

Furthermore, Jäger (2018) considers the handling of language contact (and borrowings, more specifically) an "unsolved problem for computational historical linguistics." We address this research problem and propose an automatic method for identifying the type of relationship between words (cognates or borrowings/loanwords). We show that orthographic features have discriminative power and we analyze the relevance of several underlying linguistic factors in the classification task. We run experiments on four pairs of languages: Romanian–Italian, Romanian–Spanish, Romanian–Portuguese, and Romanian–Turkish.

Our second goal is to automatically produce related words. We address the following three sub-problems: reconstructing proto-words, producing modern word forms, and producing cognates.

We begin with the reconstruction of proto-words. Given words in modern languages, the task is to automatically reconstruct the proto-words from which the modern words evolved. We address this problem in two steps. For the first step, given cognate pairs in multiple modern languages and their common ancestors, we propose a method based on sequence labeling for reconstructing proto-words that we apply on each modern language individually. For the second step, we introduce an ensemble system to use information from sister languages for reconstructing their common proto-words. We run experiments for reconstructing proto-words on three data sets of cognates in

Romance languages. We use cognate sets in Romanian, French, Italian, Spanish, and Portuguese, along with their common Latin ancestors.

Then, we address the production of modern word forms. We investigate word derivation from a donor language into a recipient language. We experiment with Romanian as a recipient language and we investigate borrowings from more than 20 donor languages. We further evaluate how well our approach models the form of foreign words that have been borrowed by Romanian, and which donor language is better modeled by our method.

Finally, for the production of cognates, we investigate whether, for a given pair of languages, having one word from a cognate pair, we can automatically determine the form of its cognate. We also conduct a comparison between recipient languages: Given a donor language whose words were borrowed in multiple recipient languages, we compare the performance of the system for each recipient language. We first use cognates between Romanian and five languages: Spanish, Italian, Turkish, Portuguese, and English. Further, we take into account the common ancestor of the cognate pairs and investigate in which language the production is better. Our experiments revolve around the Romance languages (Romanian, Italian, French, Spanish, Portuguese), but we also work with languages from other language families.

The methods that we propose can use either the orthographic or the phonetic form of the words. We use the orthographic form because this is what is available for most data sets. Moreover, we build on the idea that orthographic changes represent sound correspondences to a fairly large extent (Delmestri and Cristianini 2010). Even if the phonetic representations are largely used in identifying cognates, the orthographic representations have led to good performance as well, even on noisy data (Mackay and Kondrak 2005; Delmestri and Cristianini 2012). For one of the data sets used in our experiments, we have both the orthographic and the phonetic forms available, and we obtain very similar performance for the two cases.

3. Related Work

In a natural way, one of the most investigated problems in historical linguistics is to determine whether similar words are related or not (Kondrak 2002).

3.1 Identification of Related Words

Most studies in this area focus on automatically identifying pairs of cognates. There are three important aspects widely investigated in the task of identifying cognates: semantic, phonetic, and orthographic similarity. They were utilized both individually (Simard, Foster, and Isabelle 1992; Church 1993; Inkpen, Frunza, and Kondrak 2005) and combined (Kondrak 2004; Steiner, Stadler, and Cysouw 2011) in order to detect pairs of cognates across languages. For determining semantic similarity, external lexical resources, such as WordNet (Fellbaum 1998), might be necessary. For measuring phonetic and orthographic proximity of cognate candidates, string similarity metrics can be applied, using the phonetic or orthographic word forms as input. Various measures were investigated and compared (Inkpen, Frunza, and Kondrak 2005; Hall and Klein 2010); the edit distance (Levenshtein 1965), the XDice metric (Brew and McKelvie 1996), and the longest common subsequence ratio (Melamed 1995) are among the most frequently used metrics in this field. Gomes and Pereira Lopes (2011) proposed SpSim, a more complex method for computing the similarity of cognate pairs which tolerates learned transitions between words.

Algorithms for string alignment were successfully used for identifying cognates based on both their forms, orthographic and phonetic. Delmestri and Cristianini (2010) used basic sequence alignment algorithms (Needleman and Wunsch 1970; Smith and Waterman 1981; Gotoh 1982) to obtain orthographic alignment scores for cognate candidates. Kondrak (2000) developed the ALINE system, which aligns words' phonetic transcriptions based on multiple phonetic features and computes similarity scores using dynamic programming. List (2012) proposed a framework for automatic detection of cognate pairs, LexStat, which combines different approaches to sequence comparison and alignment derived from those used in historical linguistics and evolutionary biology.

The changes undergone by words when entering from one language into another and the transformation rules they follow have been successfully used in various approaches to identifying cognates (Koehn and Knight 2000; Mulloni and Pekar 2006; Navlea and Todirascu 2011). More recent approaches used neural networks (Rama 2016) and dictionary definitions (St Arnaud, Beck, and Kondrak 2017) to identify cognates reliably. Minett and Wang (2003) focused on identifying borrowings within a family of genetically related languages and proposed, to this end, a distance-based and a character-based technique. Minett and Wang (2005) addressed the problem of identifying language contact, building on the idea that borrowings bias the lexical similarities among genetically related languages. Tsvetkov, Ammar, and Dyer (2015) developed a model based on universal constraints from Optimality Theory to identify plausible donor-loan word pairs in contact languages.

According to the regularity principle, the distinction between cognates and borrowings benefits from the regular sound changes that generate regular phoneme correspondences in cognates (Kondrak 2002). In turn, sound correspondences are represented, to a certain extent, by alphabetic character correspondences (Delmestri and Cristianini 2010).

3.2 Production of Related Words

Kondrak (2002) drew attention to two interesting and challenging research problems in diachronic linguistics: historical derivation and comparative reconstruction. Historical derivation consists of deriving the modern forms of the words from the old ones. Comparative reconstruction is the opposite process, in which the old forms of the words are reconstructed from the modern ones.

Researchers have been continuously interested in language derivation (Pagel et al. 2013). The first attempts to address this problem focused on regular sound correspondences to construct modern forms of the words, given a proto-language, or vice versa. Some of the early studies on partially automating proto-language reconstruction belong to Covington (1998) (investigating multiple alignment for historical comparison), and Kondrak (2002) (proposing, among others, methods for cognate alignment and identification). Most of the previous approaches to producing related words relied on phonetic transcriptions (Eastlack 1977; Hartman 1981; Hewson 1974). They built on the idea that, given the phonological context, sound changes follow certain regularities across the entire vocabulary of a language. The proposed methods (Hewson 1974; Eastlack 1977; Hartman 1981) required a list of known sound correspondences as input, collected from dictionaries or published studies.

More recent approaches addressed the complete automation of the reconstruction process. Oakes (2000) proposed two systems (Jakarta and Prague) that, combined, cover the steps of the comparative method for proto-language reconstruction (discovering

regular sound changes, statistically evaluating the identified sound changes, using them to verify real word pairs, and proposing rules to infer the ancestor words from their descendants). Another probabilistic approach belongs to Hall and Klein (2010), who obtained an average edit distance of 3.8 on reconstructing proto-words using automatically determined cognate sets, using the data set of Romance languages proposed by Bouchard-Côté, Griffiths, and Klein (2009). With an average word length of 7.4, as reported by the authors, this means that, on average, the words had about half of the letters correctly determined. Bouchard-Côté et al. (2013) used probabilistic models to trace language change in the Austronesian languages, based on a given phylogenetic tree. Other probabilistic approaches to producing related words belong to Bouchard-Côté et al. (2007), Bouchard-Côté, Griffiths, and Klein (2009), and Hall and Klein (2010).

Aligning the related words to extract orthographic changes from one language to another has proven very effective when applied to both the orthographic (Gomes and Pereira Lopes 2011) and the phonetic form of the words (Kondrak 2000). The orthographic changes have also been used for producing cognates, which is closely related to the task of identifying cognates, but has not yet been as intensively studied. Whereas the purpose of identifying cognates is to determine whether two given words form a cognate pair, the aim of producing cognates is, given a word in a source language, to automatically produce its cognate pair in a target language. Beinborn, Zesch, and Gurevych (2013) proposed a method for the production of cognates relying on statistical character-based machine translation and learning orthographic production patterns, and Mulloni (2007) introduced an algorithm based on edit distance alignment and the identification of orthographic cues when words enter a new language.

One of the best approaches to reconstructing proto-words (Bouchard-Côté et al. 2013) relies on an analogy to reconstructing the genealogy of the species from genetic sequences in biology. This approach requires an existing phylogenetic tree and the phonetic transcripts of the words, to infer the ancient word forms based on probability estimates for all the possible sound changes on each branch of the tree.

4. A Dictionary-Based Approach to Building a Data Set of Related Words

In this section, we propose an algorithm for extracting cognates from electronic dictionaries that contain etymological information (Ciobanu and Dinu 2014c). After we obtain a data set of related words from dictionaries, we develop automatic methods, based on machine learning, for identifying and producing related words.

Considering a set of words in a given language L_1, to identify the cognate pairs between L_1 and a related language L_2 we apply the following strategy: First, we determine the etymologies of the given words. Then, we translate in L_2 all words without L_2 etymology. We consider cognate candidates the pairs 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, their etymologies and etymons (their source words from foreign languages). If they match, we identify the words as being cognates. We assume that etymons match even when they are different inflected forms of the same word. For example, the Romanian noun *apostrof* (*apostrophe*) has the Latin etymon *apostrophus*, which is the nominative form, and its translation in Italian, *appostrofo*, has the Latin etymon *apostrophum*, which is the accusative form. Similarly, the Romanian verb *admira* (*to admire*) has the Latin etymon *admirare*, which is the active voice (*to admire*), and its translation in Italian, *ammirare*, has the Latin etymon *admirari*, which is the passive infinitive (*to be admired*). We relax our etymon-matching rule and we identify pairs such as *apostrof-appostrofo*

and *admira-ammirare* as being cognates. This is actually a simplified, ad hoc stemming, or grouping the words with the same root. Stemming is not applicable to all languages, but it is generally accepted that the Indo-European languages, on which we run this experiment, are suited for stemming.

Our solution for addressing the task of identifying cognates answers the question raised by Swadesh (1954): "Given a small collection of likely-looking cognates, how can one definitely determine whether they are really the residue of common origin and not the workings of pure chance or some other factor?," as we limit the analysis only to words that share a common etymology—that is, words that are known to be related. For example, for the Romanian word *victorie*, Romanian dictionaries report a Latin etymology and the etymon *victoria*. Because this word does not have Italian etymology, we assume it might have a cognate pair in Italian. Consequently, we translate it in Italian, obtaining the word *vittoria*. We consider the words *victorie* and *vittoria* cognate candidates. Using an Italian dictionary, we identify, for this word, a Latin etymology and the etymon *victoria*. We compare the etymologies and the etymons for the Romanian word and its translation in Italian and, as they match, having a common ancestor (Latin) and the same etymon (*victoria*), we identify them as a cognate pair. Our method for identifying cognate pairs and word-etymon pairs is represented in Figure 1.

We investigate cognate pairs for Romanian and five other languages: French, Italian, Spanish, Portuguese, and Turkish. The first four in our list are Romance languages, and our intuition is that there are numerous words in these languages that share a common ancestor with Romanian words. As for Turkish, we decided to investigate the cognate pairs for this language because many French words were imported in both Romanian and Turkish in the nineteenth century, and we expect to find a large number of Romanian–Turkish cognate pairs with common French ancestors, which could provide a deeper insight into the lexical similarity of the two languages. The ideal situation is to use machine-readable dictionaries for all languages, but we are restricted in our investigation by the available resources.

For determining the Romanian words' etymologies, we use the DexOnline[3] machine readable dictionary. For Italian,[4] French,[5] Spanish,[6] Portuguese,[7] and Turkish[8] we extract relevant etymology-related information from online dictionaries. We use regular expressions to extract etymologies and etymons for foreign words. We manually translate Romanian words using Google Translate's first results.[9] We only needed word-level translations (not sentence-level translations). We made this choice because Google Translate is publicly available and widely used. Overall, we had to translate over 250,000 words for all the languages that we investigated. Our system does not rely on Google Translate; it can be used with other translation tools as well.

In order to evaluate our automatic method for extracting etymology-related information and for detecting related words, we randomly excerpt 500 words for each of the considered languages (Romanian, French, Italian, Spanish, Portuguese, and Turkish) and we manually determine their etymologies. Then, we compare these results with the automatically obtained etymologies and compute the accuracy for etymology extraction

3 http://dexonline.ro.
4 http://www.sapere.it/sapere/dizionari.
5 http://www.cnrtl.fr.
6 http://lema.rae.es/drae.
7 http://www.infopedia.pt/lingua-portuguesa.
8 http://www.nisanyansozluk.com.
9 http://translate.google.com.

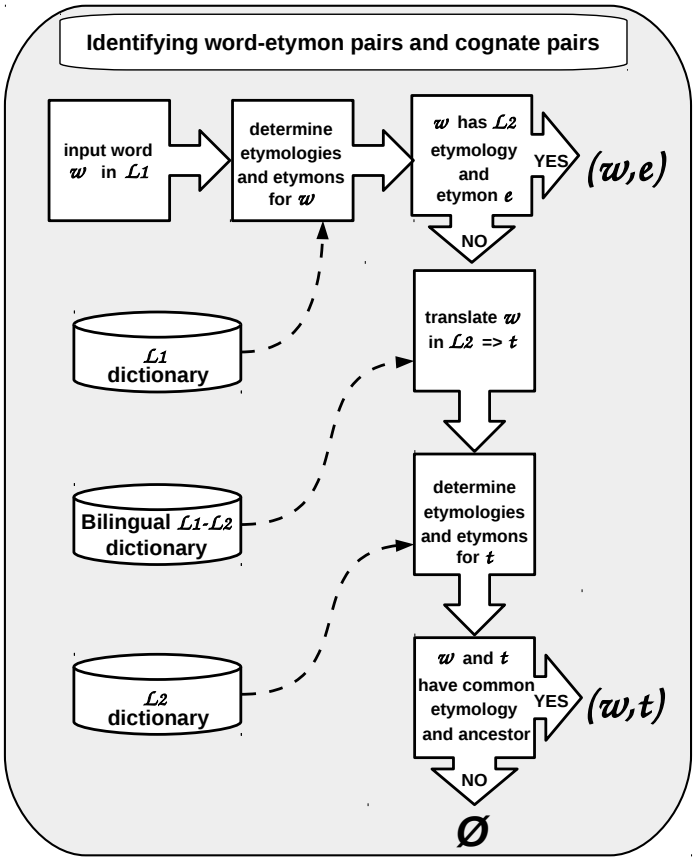

Figure 1
Identifying word-etymon pairs and cognate pairs.

for each language. We obtain the following results: 95.4% accuracy for Romanian, 98.0% for Italian, 96.6% for French, 98.2% for Spanish, 99.8% for Portuguese, and 99.6% for Turkish.

In Table 1 we report the number of Romanian words having an etymon or a cognate pair in each of the five considered languages. We account only for lexemes, leaving inflected form aside. Therefore, we consider 136,733 words in our investigation. Some of these words have cognate pairs or etymons in more than one language. A total of 4,124

Table 1
Statistics for the Romanian lexicon, regarding the number of cognates and word-etymon pairs extracted for each Romance language.

Language	no. words	no. etymons	no. cognates
French	53,347	52,868	479
Italian	13,377	9,874	3,503
Spanish	7,780	2,181	5,599
Portuguese	10,972	1,318	9,654
Turkish	4,608	2,307	2,301

Table 2
Statistics regarding the common ancestors of the identified cognate pairs.

Language	French	Italian	Spanish	Portuguese	Turkish
Arabic	–	10	15	13	4
English	3	57	94	195	158
French	–	547	455	1,925	1,157
German	–	16	14	10	–
Greek	–	221	–	1,366	410
Hebrew	–	–	1	–	–
Italian	1	–	143	238	–
Latin	475	2,606	4,874	5,815	572
Persian	–	1	–	2	–
Polish	–	–	–	2	–
Portuguese	–	3	–	–	–
Provençal	–	1	3	4	–
Russian	–	4	–	6	–
Spanish	–	34	–	72	–
Turkish	–	3	–	6	–
Total	**479**	**3,503**	**5,599**	**9,654**	**2,301**

Romanian words in DexOnline have an etymon or a cognate pair in all four Romance languages.

In Table 2 we provide statistics regarding the common ancestors of the Romanian words and their cognates in French, Italian, Spanish, Portuguese, and Turkish. As expected, most cognates between Romanian and related languages have Latin common ancestors; for Portuguese, we notice that a substantial number of cognates have French (20%) and Greek (15%) common ancestors. For Turkish, most cognates have French common ancestors as well. In the nineteenth century, numerous French words entered the Romanian lexicon. Therefore, a significant number of words are reported in the Romanian dictionaries as inherited from French. This is why the number of Romanian–French cognates is much lower than the number of words with French etymons.

5. Identification of Related Words

Our goal is to automatically identify the relationship between words, focusing on cognates and borrowings. More specifically, we propose a methodology for identifying cognates and for discriminating between cognates and borrowings. We use an orthographic alignment method and we use aligned subsequences as features for machine learning classification algorithms, in order to infer rules for linguistic changes undergone by words when entering new languages and to identify if and how the words are related.

First, we address the task of identifying cognates. That is, given a pair of words (u, v), we determine whether they are cognates or not. We apply our method on a subset of the automatically extracted data set of cognates that we previously developed for Romanian and four related Romance languages: Italian, French, Spanish, and Portuguese.

Second, we investigate the task of discriminating between cognates and borrowings. That is, given a pair of words (u, v), we determine whether they are cognates or v is the etymon of u. We run experiments on four pairs of languages: Romanian–Italian, Romanian–Spanish, Romanian–Portuguese, and Romanian–Turkish.

5.1 Methodology

Our methodology for identifying related words (cognates and borrowings) is described by the following workflow:

(1) Aligning the pairs of related words using a string alignment algorithm;

(2) Extracting features from the aligned words;

(3) Using machine learning classification algorithms to discriminate between the classes (cognates vs. non-cognates, or cognates vs. borrowings).

We utilize orthographic alignment for identifying pairs of cognates, not only to compute similarity scores, as was previously done, but to use aligned subsequences as features for machine learning algorithms. Our intuition is that inferring language-specific rules for aligning words will lead to better performance in the task of identifying cognates.

String Alignment. To align pairs of words, we use the Needleman and Wunsch (1970) global alignment algorithm. Global sequence alignment aims at determining the best alignment over the entire length of the input sequences. The algorithm guarantees finding the optimal alignment and is efficient (it uses dynamic programming).[10] Its main idea is that any partial path of the alignment along the optimal path should be the optimal path leading up to that point. Therefore, the optimal path can be determined by incremental extension of the optimal subpaths (Schuler 2002). For orthographic alignment, we consider words as input sequences and we use a very simple substitution matrix, which gives equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *è* are matched).[11]

Feature Extraction. Using the aligned pairs of words as input, we extract features around mismatches in the alignments. There are three types of mismatches, corresponding to the following operations: insertion, deletion, and substitution. For example, for

10 Every time we have multiple optimal alignments for a word pair, we take only the first one returned by the system. We use this alignment to extract features for the classification task. We also experimented with taking into account all optimal alignments (i.e., we extracted features from all of them). We did not observe significant improvements: For identifying cognates, there is a slight decrease in performance when extracting features from all optimal alignments, whereas for discriminating between cognates and borrowings there is a slight increase in performance (globally, for identification of related words, the results are, on average, 0.2% lower when extracting features from all the optimal alignments). We counted the number of word pairs from our data sets that have one optimal alignment, two optimal alignments, and so on until the maximum number. We observed that for the data set used for discriminating between cognates and non-cognates about 50% of the pairs have exactly one optimal alignment and about 80% of the pairs have less than five optimal alignments, whereas for the data set used for discriminating between cognates and borrowings about 63% of the pairs have exactly one optimal alignment and about 95% of the pairs have less than five optimal alignments.

11 We define diacritics as characters that have accents (or diacritical marks) attached. We identified in our data sets the characters whose ASCII codes are not between 'a' and 'z' or between 'A' and 'Z' and we extracted a set of diacritics to which we manually associated the corresponding letter without diacritical marks. There are currently ways of stripping diacritical marks programmatically, using open source libraries, by performing Unicode normalization followed by additional processing, but we decided to define the rules manually for higher reliability.

the Romanian word *exhaustiv* and its Italian cognate pair *esaustivo*, the alignment is as follows:

```
e x h a u s t i v -
e s - a u s t i v o
```

The first mismatch (between *x* and *s*) is caused by a substitution, the second mismatch (between *h* and -) is caused by a deletion from source language to target language, and the third mismatch (between - and *o*) is caused by an insertion from source language to target language. The features we use are character *n*-grams extracted from the alignment of the words. We ran experiments with three types of features:

(i) *n*-grams extracted around gaps in the alignment (i.e., we account only for insertions and deletions);

(ii) *n*-grams extracted around any type of mismatch in the alignment (i.e., we account for all three types of mismatches);

(iii) *n*-grams extracted from the entire alignment.

For identifying cognates, the second alternative leads to better performance, whereas for discriminating between cognates and borrowings, the third alternative leads to the highest accuracy. As for the length of the *n*-grams, we experiment with $n \in \{1, 2, 3\}$. We achieve slight improvements by combining *n*-grams with different sizes (i.e., if $n = 3$, we use 1-grams, 2-grams, and 3-grams combined). In order to provide information regarding the position of the features, we mark the beginning and the end of the word with a *$* symbol. Thus, for the above-mentioned pair of cognates, *(exhaustiv, esaustivo)*, we extract the following features around any type of mismatch in the alignment, when $n = 2$:

```
x>s ex>es xh>s-
h>- xh>s- ha>-a
->o v->vo -$>o$
```

For identical features we account only once. Therefore, because there is one feature (`xh>s-`) which occurs twice in our example, we have eight features for the pair *(exhaustiv, esaustivo)*.

Learning Algorithms. We experiment with naive Bayes and Support Vector Machines (SVMs) to learn orthographic changes and to identify the relationship between words. We put our system together using the Weka workbench (Hall et al. 2009), a suite of machine learning algorithms and tools. For SVM, we use the wrapper provided by Weka for LibSVM (Chang and Lin 2011). We use the radial basis function (RBF) kernel (Shawe-Taylor and Cristianini 2004), which can handle the case when the relation between class labels and attributes is non-linear, as it maps samples non-linearly into a higher dimensional space, making use of a kernel parameter γ.[12]

12 We also ran initial experiments with the polynomial kernel, which obtained a lower performance than the RBF kernel—on average, 4.6% lower. Thus, all SVM results reported in this article use the RBF kernel.

Evaluation Measures. To assess the performance of our method, we use the following evaluation measures: precision, recall, F-score, and accuracy.

Task Set-up. For each language pair, we split the data in two subsets, for training and testing, with a 3:1 ratio. We experiment with different values for the *n*-gram size (n $\in$ {1, 2, 3}) and we perform grid search and 3-fold cross validation over the training set in order to optimize hyperparameters c and γ. We search over $\{10^{-5}, 10^{-4}, ..., 10^{-1}, 1, 2, ..., 15\}$ for c and over $\{10^{-5}, 10^{-4}, ..., 10^{4}, 10^{5}\}$ for γ.

5.2 Automatic Identification of Cognates

In this subsection, we apply our methodology to automatically determine pairs of cognates across languages (Ciobanu and Dinu 2014b).

Words undergo various changes when entering new languages. We assume that rules for adapting foreign words to the orthographic system of the target languages might not have been very well defined in their period of early development, but they may have since become complex and probably language-specific. Detecting pairs of cognates based on etymology is useful and reliable; but for resource-poor languages, methods that require less linguistic knowledge might be necessary. The proposed method requires a list of known cognates; and for languages for which additional linguistic information is available, it can be customized to integrate historical information regarding the evolution of the language.

5.2.1 Experiments. We apply our method on an automatically extracted data set of cognates for four pairs of languages: Romanian–French, Romanian–Italian, Romanian–Spanish, and Romanian–Portuguese. We use the data set introduced in Section 4. We discard pairs of words for which the forms across languages are identical (i.e., the Romanian word *matrice* and its Italian cognate pair *matrice*, having the same form), because these pairs do not provide any orthographic changes to be learned. For each pair of languages we determine a number of non-cognate pairs equal to the number of cognate pairs. The non-cognates are also translations of the Romanian words in French, Italian, Spanish, and Portuguese, but the difference is that they do not share a common etymon with the Romanian words. Finally, we obtain 445 pairs of cognates for Romanian–French,[13] 3,477 for Romanian–Italian, 5,113 for Romanian–Spanish, and 7,858 for Romanian–Portuguese. Because we need sets of approximately equal size for comparison across languages, we keep 400 pairs of cognates and 400 pairs of non-cognates for each pair of languages. In Table 3 we report statistics regarding the length of the words in the data set, and the average edit distance between cognates and non-cognates. Given a pair of languages (L_1, L_2), the len_1 and len_2 columns represent the average word length of the words in L_1 and L_2, respectively. The edit column represents the average normalized edit distance between the words. The values are computed only on the training data, to keep the test data unseen. The difference in length between the related words shows what operations to expect when aligning the words. In Tables 4 and 5, we provide, for each pair of languages, the five most relevant 2-gram orthographic changes, determined using the χ^2 distribution implemented in

13 The number of pairs of cognates is much lower for French than for the other languages because there are numerous Romanian words that have French etymology and here we do not consider these words to be cognate candidates.

Table 3
Statistics for the data set of cognates and non-cognates.

Lang.	Cognates			Non-cognates		
	len_1	len_2	edit	len_1	len_2	edit
It–Ro	8.01	8.92	0.26	7.90	8.16	0.57
Fr–Ro	8.62	8.51	0.40	7.02	6.85	0.76
Es–Ro	7.91	8.30	0.26	7.39	7.40	0.67
Pt–Ro	8.14	8.49	0.29	8.22	8.22	0.54

Weka, and the five most frequent 2-gram orthographic changes in the cognate pairs from our data set.[14] The χ^2 test determines the dependency of two variables. For feature selection, the class label is used as target variable. The χ^2 statistics is computed for each feature, with respect to the target variable, and the features are ranked based on the computed values. The lower ranked features are those independent of the target variable (the class label). They are deemed less useful for the classification task. None of the top ranked orthographic cues occurs at the beginning of the word, while many of them occur at the end of the word. The most frequent operation in Tables 4 and 5 is substitution.

Baselines. We compare the performance of the method we propose with previous approaches for automatic detection of cognate pairs based on orthographic similarity. We use several orthographic metrics widely used in this research area: the edit distance (Levenshtein 1965), the longest common subsequence ratio (Melamed 1995), and the XDice metric (Brew and McKelvie 1996).[15] In addition, we use SpSim (Gomes and Pereira Lopes 2011), which outperformed the longest common subsequence ratio and a similarity measure based on the edit distance in previous experiments. To evaluate these metrics on our data set, we follow the strategy described in Inkpen, Frunza, and Kondrak (2005). First, we compute the pairwise distances between pairs of words for each orthographic metric individually, as a single feature.[16] In order to detect the best threshold for discriminating between cognates and non-cognates, we run a decision stump classifier (provided by Weka) on the training set for each pair of languages and for each metric. A decision stump is a decision tree classifier with only one internal node and two leaves corresponding to our two class labels. Using the best threshold value selected for each metric and pair of languages (and using accuracy for optimization), we further classify the pairs of words in our test sets as cognates or non-cognates.

5.2.2 Results and Analysis. In Table 6 we report the results for automatic identification of cognates using orthographic alignment. We report the *n*-gram values for which the best

14 For brevity, we use in the tables the ISO 639-1 codes for language abbreviation. We denote pairs of languages by the target language, given the fact that Romanian is always the source language in our experiments.

15 We use normalized similarity metrics for all experiments. For the edit distance, we subtract the normalized value from 1 in order to obtain similarity.

16 SpSim cannot be computed directly, as the other metrics, so we introduce an additional step in which we use one third of the training set (only cognates are needed) to learn orthographic changes. In order to maintain a stratified data set, we discard an equal number of non-cognates in the training set and then we compute the distances for the rest of the training set and for the test set. We use the remainder of the initial training set for the next step of the procedure.

Table 4
The most relevant orthographic cues for each pair of languages determined on the entire data sets using the χ^2 attribute evaluation method implemented in Weka.

Rank	It–Ro	Fr–Ro	Es–Ro	Pt–Ro
1	iu > io	un > on	−$ > o$	ie > ão
2	un > on	ne > n−	ţi > ci	aţ > aç
3	l− > le	iu > io	− > ón	ţi > çã
4	t$ > − $	ţi > ti	ie > ió	i$ > −$
5	−$ > e$	e$ > − $	at > ad	ă$ > a$

Table 5
The most frequent orthographic cues for each pair of languages determined on the cognate lists using the raw frequencies.

Rank	It–Ro	Fr–Ro	Es–Ro	Pt–Ro
1	−$ > e$	e$ > −$	$ > o$	−$ > o$
2	−$ > o$	un > on	e$ > −$	ă$ > a$
3	ă$ > a$	ne > n−	ţi > ci	e$ > −$
4	− > re	iu > io	ă$ > a$	−$ > r$
5	ţi > zi	ţi > ti	at > ad	−$ > a$

results are obtained and the hyperparameters for SVM, c, and γ. The best results are obtained for French and Spanish, and the lowest accuracy is obtained for Portuguese. The SVM produces better results for all considered languages except Portuguese, where the accuracy is equal. For Portuguese, both naive Bayes and SVM misclassify more non-cognates as cognates than vice versa. A possible explanation might be the occurrence, in the data set, of more remotely related words, which are not labeled as cognates. To test this assumption, we analyzed the errors of the system and observed that around 38% of the pairs misclassified as cognates are actually more remotely related words.

In Table 7 we report the results of previous methods for identifying cognates, for comparison. We observe that our method outperforms the orthographic metrics considered as individual features.

5.3 Cognates vs. Borrowings

In this subsection, we address the task of automatically distinguishing between borrowings and cognates (Ciobanu and Dinu 2015). Given a pair of words, the task is to determine whether one is a historical descendant of the other, or whether they both share a common ancestor. For this experiment, we know a priori that the words are related. Further, we extend the research problem adding a new class for unrelated words. In other words, we have a classification task with three classes: cognates, borrowings, and unrelated words.

5.3.1 Experiments. We apply our method on four pairs of languages extracted from the data set introduced in Section 4: Italian–Romanian, Portuguese–Romanian, Spanish–Romanian, and Turkish–Romanian. For the first three pairs of languages, which include sister languages, most cognate pairs have a Latin common ancestor, while for the fourth pair, which includes languages belonging to different families (Romance and Turkic),

Table 6
Results for automatic detection of cognates using orthographic alignment. We report the precision (P), recall (R), F-score (F), and accuracy (A) obtained on the test sets and the optimal n-gram values. For SVM we also report the optimal hyperparameters c and γ obtained during cross-validation on the training sets.

Lang.	Naive Bayes					SVM						
	P	**R**	**F**	**A**	n	**P**	**R**	**F**	**A**	n	c	γ
It–Ro	72.7	93.0	81.6	79.0	1	76.0	92.0	83.2	81.5	1	1	0.10
Fr–Ro	81.3	91.0	85.9	82.0	2	84.9	89.0	86.9	87.0	2	10	0.01
Es–Ro	79.3	92.0	85.2	84.0	1	85.4	88.0	86.7	86.5	2	4	0.01
Pt–Ro	67.7	88.0	76.5	73.0	2	70.9	78.0	74.3	73.0	2	10	0.01

Table 7
Comparison with previous methods for automatic detection of cognate pairs based on orthography. We report the precision (P), recall (R), F-score (F), and accuracy (A) obtained on the test sets and the optimal threshold t for discriminating between cognates and non-cognates.

Lang.	P	R	F	A	t
It–Ro	67.4	97.0	79.5	75.0	0.43
Fr–Ro	76.2	93.0	83.8	82.0	0.30
Es–Ro	77.1	91.0	83.5	82.0	0.56
Pt–Ro	62.3	99.0	76.5	69.5	0.34

(a) EDIT

Lang.	P	R	F	A	t
It–Ro	68.9	91.0	78.4	75.0	0.51
Fr–Ro	76.9	90.0	82.9	81.5	0.42
Es–Ro	72.4	97.0	82.9	80.0	0.47
Pt–Ro	59.3	99.0	74.2	65.5	0.34

(b) LCSR

Lang.	P	R	F	A	t
It–Ro	66.2	98.0	79.0	74.0	0.21
Fr–Ro	77.5	79.0	78.2	78.0	0.26
Es–Ro	72.3	99.0	83.6	80.5	0.19
Pt–Ro	57.9	99.0	73.1	63.5	0.10

(c) XDICE

Lang.	P	R	F	A	t
It–Ro	66.7	98.0	79.4	74.5	0.44
Fr–Ro	86.5	83.0	84.7	85.0	0.59
Es–Ro	81.8	90.0	85.7	85.0	0.64
Pt–Ro	62.2	97.0	75.8	69.0	0.39

(d) SPSIM

most of the cognate pairs have a common French etymology, and date back to the end of the nineteenth century, when both Romanian and Turkish borrowed massively from French. In Table 8 we provide examples of borrowings and cognates.

The data set contains borrowings[17] and cognates that share a common ancestor. We use a stratified data set of 2,600 pairs of related words for each pair of languages (that is, we have 1,300 pairs of cognates and 1,300 pairs of borrowings). In Table 9 we provide an initial analysis of our data set. We report statistics regarding the length of the words and the edit distance between them. Romanian words are almost in all situations shorter, on average, than their pairs. For Turkish–Romanian, len_1 is higher than len_2, so we expect more deletions for this pair of languages. The edit columns show how much words vary from one language to another based on their relationship (cognates or borrowings). For Italian–Romanian, both distances are small (0.26 and 0.29), as opposed to the other languages, where there is a more significant difference between the two (e.g., 0.26 and 0.52 for Spanish–Romanian). The small difference for Italian–Romanian might make the discrimination between the two classes more difficult.

17 Romanian is always the recipient language in our data set (i.e., the language that borrowed the words).

Table 8
Examples of borrowings and cognates for Romanian. For cognates we also report the common ancestor.

Lang.	Borrowings			Cognates		
It–Ro	baletto	→	balet	vittoria - victorie		↑ victoria
	(ballet)			(victory)		(Latin)
Pt–Ro	selva	→	selvă	instinto - instinct		↑ instinctus
	(selva)			(instinct)		(Latin)
Es–Ro	machete	→	macetă	castillo - castel		↑ castellum
	(machete)			(castle)		(Latin)
Tr–Ro	tütün	→	tutun	aranjman - aranjament		↑ arrangement
	(tobacco)			(arrangement)		(French)

Table 9
Statistics for the data set of cognates and borrowings.

Lang.	Cognates			Borrowings		
	len_1	len_2	edit	len_1	len_2	edit
It–Ro	7.95	8.78	0.26	7.58	8.41	0.29
Pt–Ro	7.99	8.35	0.28	5.35	5.42	0.52
Es–Ro	7.91	8.33	0.26	5.78	6.14	0.52
Tr–Ro	7.35	6.88	0.31	6.49	6.09	0.44

Baselines. Given the initial analysis presented above, we hypothesize that the distance between the words might be indicative of the type of relationship between them. Previous studies (Inkpen, Frunza, and Kondrak 2005; Gomes and Pereira Lopes 2011) show that related and non-related words can be distinguished based on the distance between them, but a finer-grained task, such as determining the type of relationship between words, is probably more subtle. We compare our method with two baselines:

- A baseline that assigns a label based on the normalized edit distance between the words: given a test instance pair $word_1$ - $word_2$, we subtract the average normalized edit distance between $word_1$ and $word_2$ from the average normalized edit distance of the cognate pairs and from the average normalized edit distance between the borrowings and their etymons (computed on the training set; see Table 9), and assign the label that yields a smaller difference (in absolute value). In case of equality, the label is chosen randomly.
- A decision tree classifier, following the strategy proposed by Inkpen, Frunza, and Kondrak (2005): we use the normalized edit distance as single feature, and we fit a decision tree classifier with the maximum tree depth set to 1. We perform 3-fold cross-validation in order to select the best threshold for discriminating between borrowings and cognates. Using the best threshold selected for each language, we further assign one of the two classes to the pairs of words in our test set.

Table 10
Average precision (P), recall (R), F-score (F), and accuracy (A) for automatic discrimination between cognates and borrowings.

Lang.	Baseline #1				Baseline #2			
	P	**R**	**F**	**A**	**P**	**R**	**F**	**A**
It–Ro	50.7	50.7	50.7	50.7	64.4	54.5	59.0	54.4
Pt–Ro	79.3	79.0	79.1	79.0	80.1	80.0	80.0	80.0
Es–Ro	78.6	78.4	78.5	78.5	78.6	78.5	78.5	78.4
Tr–Ro	61.1	60.6	60.8	61.3	62.5	59.8	61.1	59.8

Table 11
Average precision (P), recall (R), F-score (F), accuracy (A), and optimal n-gram size for automatic discrimination between cognates and borrowings. For SVM we also report the optimal values for c and γ.

Lang.	Naive Bayes					SVM						
	P	**R**	**F**	**A**	n	**P**	**R**	**F**	**A**	n	c	γ
It–Ro	68.6	68.2	68.4	68.1	3	69.2	69.1	69.1	69.0	3	10	0.10
Pt–Ro	92.6	91.7	92.1	91.6	3	90.1	90.0	90.0	90.0	3	3	0.10
Es–Ro	85.3	84.5	84.9	84.4	3	85.7	85.5	85.6	85.5	2	2	0.10
Tr–Ro	89.7	89.4	89.5	89.3	3	90.3	90.2	90.2	90.1	3	6	0.01

5.3.2 Results and Analysis. Table 10 and Table 11 show the results for automatic discrimination between cognates and borrowings. The two baselines produce comparable results. For all pairs of languages, our method significantly outperforms the baselines (99% confidence level)[18] with values between 7% and 29% for the accuracy, suggesting that the n-grams extracted from the alignment of the words are better indicators of the type of relationship than the edit distance between them. SVM obtains, in most cases, better results than naive Bayes. The best results are obtained for Turkish–Romanian, with an accuracy of 90.1, followed by Portuguese–Romanian with 90.0 and Spanish–Romanian with 85.5 (for Portuguese–Romanian, naive Bayes obtains a slightly better result than SVM, with an accuracy of 91.6). These results show that, for these pairs of languages, the orthographic cues are different with regard to the relationship between words. For Italian–Romanian we obtain the lowest accuracy, 69.0.

In this experiment, we know beforehand that there is a relationship between words, and our aim is to identify the type of relationship. However, in many situations this kind of a priori information is not available. In a real scenario, we would have to either add an intermediary classifier for discriminating between related and unrelated words, or to discriminate between three classes: cognates, borrowings, and unrelated. We augment our data set with unrelated words (determined based on their etymology), building a stratified data set annotated with three classes, and we repeat the previous experiment. The performance decreases, but the results are still significantly better than chance (99% confidence level). We obtained the following accuracy values on the test sets, when using the SVM classifier: Italian–Romanian 63.8, Portuguese–Romanian 77.6, Spanish–Romanian 74.0, and Turkish–Romanian 86.0. For Italian, borrowings turned out to be

18 All the statistical significance tests reported in here are performed on 1,000 iterations of paired bootstrap resampling (Koehn 2004).

Table 12
Average precision (P), recall (R), F-score (F), accuracy (A), and optimal *n*-gram size for automatic discrimination between cognates and borrowings using various linguistic factors as additional features.

Lang.	Naive Bayes					SVM						
	P	R	F	A	*n*	P	R	F	A	*n*	*c*	γ
It–Ro	68.7	68.2	68.4	68.2	3	68.1	68.1	68.1	68.0	3	7	0.01
Pt–Ro	91.8	91.0	91.4	91.0	3	92.3	92.3	92.3	92.2	2	7	0.01
Es–Ro	86.0	85.2	85.6	85.1	3	85.6	85.5	85.5	85.4	3	9	0.01
Tr–Ro	89.8	89.8	89.8	89.7	2	89.0	88.9	88.8	88.8	2	4	0.10
(a) Part of speech												
It–Ro	65.7	65.5	65.5	65.5	2	67.4	67.3	67.3	67.3	2	1	0.10
Pt–Ro	93.4	92.9	93.2	92.9	2	91.9	91.9	91.9	91.9	2	4	0.01
Es–Ro	86.7	86.3	86.4	86.3	2	85.6	85.6	85.6	85.5	2	10	0.10
Tr–Ro	84.9	84.7	84.8	84.7	2	88.8	88.7	88.7	88.6	1	3	0.10
(b) Hyphenization												
It–Ro	66.2	66.0	66.1	66.0	2	65.9	65.8	65.8	65.8	2	1	0.10
Pt–Ro	83.3	82.9	83.0	82.9	2	85.1	85.1	85.1	85.0	2	8	0.10
Es–Ro	82.5	81.8	82.1	81.8	3	78.8	78.8	78.8	78.7	3	10	0.01
Tr–Ro	0.80	79.7	79.8	79.6	3	78.1	78.0	78.0	78.0	3	7	0.01
(c) Consonants												
It–Ro	62.4	61.2	61.3	61.2	2	63.0	63.3	62.7	63.3	1	8	0.10
Pt–Ro	90.8	89.7	90.3	89.7	3	90.3	90.2	90.2	90.2	1	2	0.01
Es–Ro	79.3	78.8	78.8	78.7	2	80.4	80.3	80.2	80.3	3	4	0.01
Tr–Ro	85.8	85.5	85.5	85.5	2	87.1	86.8	86.7	86.7	3	5	0.01
(d) Stems												
It–Ro	66.5	66.2	66.3	66.1	3	66.8	66.7	66.7	66.7	2	2	0.10
Pt–Ro	92.4	91.7	92.0	91.6	3	91.3	91.3	91.3	91.2	2	10	0.01
Es–Ro	83.6	83.2	83.4	83.1	3	83.2	83.0	83.0	82.9	2	10	0.10
Tr–Ro	89.4	89.3	89.3	89.2	3	88.1	88.0	88.0	88.0	2	10	0.10
(e) Diacritics												

the most difficult class to identify correctly. For Turkish, the cognates were slightly more difficult to identify correctly compared to other classes, while for Portuguese and Spanish the lowest performance was obtained for unrelated words. In both cases, they were more often mistakenly identified as cognates than as borrowings.

Linguistic Factors. To gain insight into the factors with high predictive power, we perform several further experiments. The results are reported in Table 12.

- **Part of speech.** We investigate whether adding knowledge about the part of speech of the words leads to performance improvements. Verbs, nouns, adverbs, and adjectives have language-specific endings; thus we assume that part of speech (POS) might be useful when learning orthographic patterns. We obtain POS tags from the DexOnline machine-readable dictionary. We use the POS feature as an additional categorical feature for the learning algorithm. It turns out that, except for Portuguese–Romanian (accuracy 92.2), the additional POS feature does not improve the performance of our method.

- **Hyphenization.** We analyze whether the system benefits from using the hyphenated form of the words as input to the alignment algorithm. We are interested to see whether marking the boundaries between the syllables improves the alignment (and, thus, the feature extraction). The hyphen boundaries ("|") are considered as additional characters in the input strings. We obtain the hyphenization for the words in our data set from the RoSyllabiDict dictionary (Barbu 2008) for Romanian words and several available Perl modules[19] for the other languages. For Portuguese–Romanian the accuracy increases by about 2 percentage points, reaching a value of 91.9.
- **Consonants.** We examine the performance of our system when trained and tested only on the aligned *consonant skeletons* of the words (i.e., a version of the words where vowels are discarded). According to Ashby and Maidment (2005), consonants change at a slower pace than vowels across time; while the former are regarded as reference points, the latter are believed to carry less information useful for identifying the words (Gooskens, Heeringa, and Beijering 2008). The performance of the system decreases when vowels are removed (95% confidence level). We also train and test the decision tree classifier on this version of the data set, and its performance is lower in this case as well (95% confidence level), indicating that, for our task, the information carried by the vowels is helpful.
- **Stems.** We repeat the first experiment using stems as input, instead of lemmas. What we seek to understand is whether the aligned affixes are indicative of the type of relationship between words. We use the Snowball Stemmer[20] and we find that the performance decreases when stems are used instead of lemmas. Performing a χ^2 feature ranking on the features extracted from mismatches in the alignment of the related words reveals further insight into this matter: For all pairs of languages, at least one feature containing the $ character (indicating the beginning or the end of a word) is ranked among the 10 most relevant features, and over 50 are ranked among the 500 most relevant features. This suggests that prefixes and suffixes (usually removed by the stemmer) vary with the type of relationship between words.
- **Diacritics.** We explore whether removing diacritics influences the performance of the system. Many words have undergone transformations by the augmentation of language-specific diacritics when entering a new language. For this reason, we expect diacritics to play a role in the classification task. We observe that, when diacritics are removed, the accuracy on the test set is lower in almost all situations. Analyzing the ranking of the features extracted from mismatches in the alignment provides even stronger evidence in this direction: For all pairs of languages, more than a fifth of the top 500 features contain diacritics.

19 Modules Lingua::It::Hyphenate, Lingua::Pt::Hyphenate, Lingua::Es::Hyphenate, Lingua:Tr:::Hyphenate, available on the Comprehensive Perl Archive Network: `www.cpan.org`.

20 `http://snowball.tartarus.org`.

6. Production of Related Words

We present a method for producing related words based on orthography. We account for the type of relationship between words, making a clear distinction between proto-words, borrowed words, and cognates. Our goal is to achieve state-of-the-art performance in reconstructing proto-words using less resources than in previous studies—that is, without using a lexicon or a data set in the recipient language, for example. We aim at providing a tool for linguists to use in their research.

6.1 Methodology

In this section we introduce a new technique for the production of related words. Our goal is to develop a tool that would provide support in historical linguistics, where the produced words would be further analyzed by domain experts. We propose an approach based on conditional random fields. We apply a sequence labeling method that predicts the form of the related words.

Conditional Random Fields. From the alignment of related words in the training set, the system learns orthographic patterns for the changes in spelling between the source and the target language. The method that we utilize is based on sequence labeling, an approach that has been proven useful in generating transliterations (Ganesh et al. 2008; Ammar, Dyer, and Smith 2012).

In our case, the words are the sequences, and their characters are the tokens. Our purpose is to obtain, for each input word, a sequence of characters that compose its related word. To this end, we use conditional random fields (CRFs) (Lafferty, McCallum, and Pereira 2001). As features for the CRF system, we use character n-grams from the input words, extracted from a fixed window w around the current token.

Pairwise Sequence Alignment. To align pairs of words, we use the Needleman and Wunsch (1970) global alignment algorithm.[21] For example, for the Romanian word *frumos* (meaning *beautiful*) and its Latin ancestor *formosus*, the alignment is as follows:

```
f - r u m o s - -
f o r - m o s u s
```

For each character in the source word (after the alignment), the associated label is the character that occurs on the same position in the target word. In the case of insertions, we add the new character to the previous label, because there is no input character in the source language to which we could associate the inserted character as label. We account for affixes separately: For each input word, we add two more characters B and E, marking the beginning and the end of the word. The characters that are inserted in the target word at the beginning or at the end of the word are associated with these

21 In a preliminary step, we experimented with two alignment methods: the method based on profile hidden Markov models proposed by Bhargava and Kondrak (2009) and the Needleman and Wunsch (1970) global alignment algorithm. Because the alignment is only a pre-processing step for our task, we evaluate the alignment methods by the downstream results, that is, the accuracy obtained by the CRF system when using one or the other alignment method. We observed that the results obtained with the two alignment methods were very similar, slightly better for the latter. Thus, we report the results obtained with the Needleman Wunsch alignment algorithm, following the methodology presented in Section 5.1.

special characters. In order to reduce the number of labels, we replace the label with * for input tokens that are identical to their labels. Thus, for the previous example, the labels are as follows:[22]

```
B f r u m o s  E
↓ ↓ ↓ ↓ ↓ ↓ ↓  ↓
* fo * - * * *  usE
```

Evaluation Measures. Following previous work in this area (Bouchard-Côté et al. 2007; Beinborn, Zesch, and Gurevych 2013), we use the following evaluation measures to assess the performance of our method:

- **Average edit distance.** To assess how close the productions are to the correct form of their related words, we report the edit distance between the produced words and the gold standard. We report both the un-normalized and the normalized edit distance. For normalization in the [0,1] interval, we divide the edit distance by the length of the longest string.
- **Coverage.** Also known as **top *n* accuracy**, the coverage is a relaxed metric that computes the percentage of input words for which the n-best output list contains the correct proto-word (the gold standard). We use $n \in \{1, 5, 10\}$. The practical importance of analyzing the top n results is that we offer a filter to narrow down the possible forms of the output words to a low-dimensional list, which linguists can analyze, aiming to identify the correct form of the proto-word. Note that the coverage for $n = 1$ is the well-known measure accuracy.
- **Mean reciprocal rank.** The mean reciprocal rank is an evaluation measure that applies to systems producing an ordered output list for each input instance. Given an input word, the higher the position of its correct proto-word in the output list, the higher the mean reciprocal rank value:

$$MRR(w_i) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{rank_i} \tag{1}$$

  where m is the number of input instances, and $rank_i$ is the position of w_i's proto-word in the output list. If w_i's correct proto-word is not in the output list, we consider the reciprocal rank 0.

Task Set-up. We split each data set in subsets for train, development, and test (3:1:1 ratio). For inferring the form of the related words, we use the CRF implementation

22 We also investigated several methods for further improving the results, but we obtained no significant improvements. First, we investigated a measure of reranking based on occurrence probabilities for the n-grams in the produced sequences adapted from language modeling (Zhang, Hildebrand, and Vogel 2006), using only the training data to obtain probabilities. This approach did not produce good results, most probably because of insufficient data. Second, we split the training data set based on the part of speech of the words. Based on the intuition that certain orthographic patterns are specific to certain parts of speech, we investigated whether training separate models would produce better results. Although for some parts of speech the results improved, overall the performance of the model was lower than that of the CRF model followed by maximum entropy reranking.

provided by the Mallet toolkit (McCallum 2002). For parameter tuning, we perform a grid search for the number of iterations in $\{1, 5, 10, 25, 50, 100\}$ and for the size of the window w in $\{1, 2, 3, 4, 5\}$.

6.2 Reconstruction of Proto-words

We address the problem of reconstructing proto-words in two steps (Ciobanu and Dinu 2018). First, given cognate pairs in multiple modern languages and their common ancestors, we apply a method based on sequence labeling for reconstructing proto-words. We apply the method on each modern language individually. Second, we propose several ensemble methods for combining information from multiple systems, with the purpose of joining the best productions from all modern languages. Through experiments, we show that our best ensemble system outperforms previous results Bouchard-Côté et al. 2007. The novelty of our approach is enhancing the sequence alignment system with an additional step of reranking. We also compute the results of an oracle, which shows the potential of this approach. The proposed methods have the advantage of not requiring phonetic transcripts and other data besides the training word pairs (such as a corpus in the target language, as some of the existing methods require); external information regarding language evolution is difficult to obtain for some languages, and this method can be applied on low-resourced and endangered languages. Moreover, as opposed to previous methods, our system is able to reconstruct proto-words even from incomplete cognate sets (cognate pairs in multiple modern languages descending from a common proto-language).

6.2.1 Ensembles. Ensembles of classifiers combine the results of multiple classifiers, in order to improve performance. In our case, the classifiers use different input data: We train a classifier for each modern language and combine their results, in order to obtain Latin proto-words with higher accuracy. Our goal is to improve the performance of the system by taking advantage of the information provided by all considered languages.

Each classifier produces, for each input word, an n-best list of possible proto-words. To combine the outputs of the classifiers, we propose four fusion methods based on the ranks in the n-best lists and the probability estimates provided by the individual classifiers for each possible production.

Given a cognate set, we combine the n-best lists previously obtained to compute a joint n-best list that leverages information from all modern languages. Our methodology is illustrated in Figure 2.

Fusion by Rank. We compute an average rank for each production from the n-best lists, as the average rank from all modern languages. We favor words that occur on the first position in multiple n-best lists. In other words, for a given n-best list, we associate weights in a de Borda (1781) sense to all the words from the list: we give weight n to the word produced with the highest confidence, $n - 1$ to the second one, and so on, until weight *1* to the n-th produced word. For an n-best list L_i and a word u, we denote by $w(u_i)$ the weight of word u in L_i; if u is not a word from list L_i, then $w(u_i) = 0$. Given k n-best lists (one for each modern language) and a produced word u, we define the rank weight of u over the k n-best lists as:

$$w_r(u) = (1/k) * \sum_{i=1}^{k} w(u_i) \tag{2}$$

Fusion by Rank and Accuracy. Starting from the previous fusion method, we also take into account the training accuracy for each language. We give more importance to the *vote* of the languages that obtained a better performance, multiplying the weight by the training accuracy. In other words, for each n-best list L_i, let $\pi(i)$ be the training accuracy for language i. Given k n-best lists (one for each modern language) and a produced word u, we define the rank-accuracy weight of u over the k n-best lists as:

$$w_{ra}(u) = (1/k) * \sum_{i=1}^{k} w(u_i)\pi(i) \tag{3}$$

Fusion by Weight. We compute an average confidence score for each production from the n-best lists, using the confidence score reported by the sequence labeling system. We rerank the productions based on the average confidence score. Given k n-best lists (one for each modern language), a produced word u, and $w(u_i)$ the confidence score of

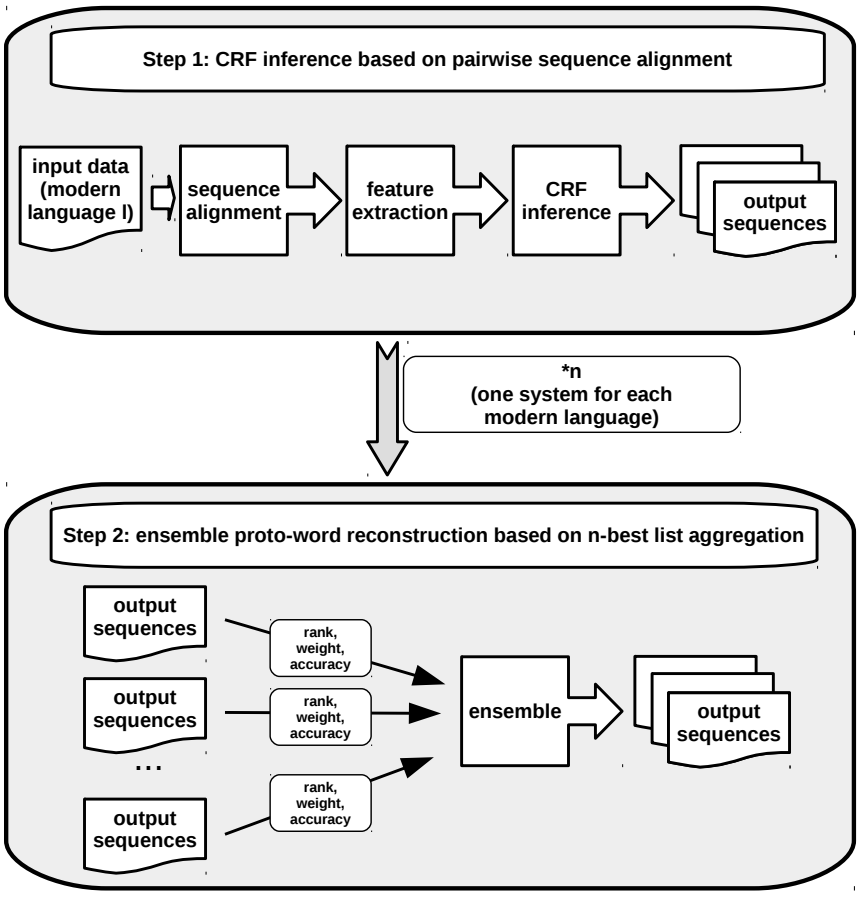

Figure 2
Methodology for reconstructing proto-words.

the sequence labeling system in list L_i, we define the confidence weight of u over the k n-best lists as:

$$w_c(u) = (1/k) * \sum_{i=1}^{k} w(u_i) \tag{4}$$

This is similar to the first fusion method, but uses the sequence labeling system's weights instead of the weights obtained from the ranking in the n-best lists.

Fusion by Weight and Accuracy. Starting from the previous fusion method, we also take into account the training accuracy for each language. We give more importance to the *vote* of the languages that obtained a better performance, multiplying the score by the training accuracy. Given k n-best lists (one for each modern language), a produced word u, and $w(u_i)$ the confidence score of the sequence labeling system in list L_i, we define the confidence-accuracy weight of u over the k n-best lists as:

$$w_{ca}(u) = (1/k) * \sum_{i=1}^{k} w(u_i)\pi(i) \tag{5}$$

This is similar to the second fusion method, but uses the sequence labeling system's weights instead of the weights obtained from the ranking in the n-best lists.

The output of each ensemble is a new n-best list in which the words are sorted in descending order of their computed weights, as described in Equations (2)–(5).

Oracle. An oracle classifier is an ensemble method that produces the correct result if any of the included classifiers produces the correct result. The purpose of an oracle is to determine the upper limit of an ensemble.

6.2.2 Experiments. We use data sets of Romance languages with Latin ancestors:

Data set 1. The data set proposed by Bouchard-Côté et al. (2007). It contains 585 complete cognate sets in three Romance languages (Spanish, Portuguese, Italian) and their common Latin ancestors. It is provided in two versions: orthographic and phonetic (IPA transcriptions). This data set allows us to compare our results with a previous state-of-the-art method for reconstructing proto-words.

Data set 2. The data set proposed by Reinheimer Ripeanu (2001). It consists of 1,102 cognate sets in five Romance languages (Spanish, Italian, Portuguese, French, Romanian) and their common Latin ancestors. Note that not all of these cognate sets are complete (i.e., for some of them there are not cognates provided in all five modern languages).

Data set 3. The data set introduced in Section 4. It contains 3,218 complete cognate sets in five Romance languages (Spanish, Italian, Portuguese, French, Romanian) and their common Latin ancestors. An example of a cognate set from this data set:

vehicul (**Ro**) | véhicule (**Fr**) | veicolo (**It**) | vehículo (**Es**) | veículo (**Pt**) | vehiculum (**Lat**)

6.2.3 Results and Analysis. In Table 13 we report the results of our individual systems (one for each modern language) and the ensemble results for reconstructing proto-words. For individual experiments, Italian obtains the lowest average edit distance on all data sets.

For ensembles, we experimented with the four fusion methods described in the previous section, but report only the best performing one. Out of the four proposed fusion methods, the first two lead to similar results, which are superior to the other two. The best-performing ensemble uses the second fusion method, which assigns scores based on the rank in the *n*-best lists and the training accuracy for each individual system. We also tried applying the ensembles on language subsets (i.e., not to take all modern languages into account at once). We investigated all combinations, and in the majority of cases using all modern languages leads to the highest performance among all ensembles.

In Table 14 we show an example of our systems' output *n*-best lists. This example illustrates how the ensemble can improve on the individual classifiers, by ranking the correct production higher than all the other systems. For all data sets, we obtained performance improvements for reconstructing proto-words when we combined individual results using ensembles.

As expected, the highest performance was obtained by the oracle classifiers. The results show the high potential of the ensemble methods for reconstructing proto-words. The average edit distance of our best-performing ensemble, on Data set 3, is 1.07, meaning that, on average, the reconstructions obtained by the system are a little more than one character different from the correct proto-words. Furthermore, the correct proto-word is listed among the 5-best list productions of our system in 70% of the cases (increasing to 74% for 10-best lists). These results are encouraging, having in mind the purpose of our system: to be a tool to be used by linguists (not to substitute the work of the experts).

Overall, we notice that the results are significantly better for Data sets 2 and 3 than for Data set 1. One possible explanation is the nature of the data set: whereas Data sets 2 and 3 are built based on the etymologies of the words (that is, the genetic relationships are taken into account), for Data set 1 the cognacy decisions have been made based on an edit distance threshold between the words (Bouchard-Côté et al. 2007).

To test this assumption, we ran an additional experiment, training the system on a subset of Data set 3 having the same size as Data set 1. The performance was close to that reported for Data set 3 in Table 13, confirming that it is the nature of the data set rather than its size that influences the results.

Another interesting case is that of the cognate sets that could be identified in the modern Romance languages, but for which it was not possible to trace any evidence of their Latin ancestor. In this situation, the Latin form has been artificially generated. We applied our best-performing ensemble system on four such Latin words: *oblitare* (Brodsky 2009), *fortia* (Alkire and Rosen 2010), *barra*, and *blancus*. The corresponding cognate sets are reported in Table 15.

The system was able to reconstruct the Latin proto-word correctly in three of the four cases. In two cases, the correct proto-word was the first prediction of the system (*blancus*, *barra*), and in another case the correct Latin proto-word was on the second position in the system's *n*-best list of productions (*fortia*). For *oblitare*, the system's second production was *obliare*, only one letter different from the correct proto-word.

Error Analysis. In this subsection we perform a brief analysis of the errors of our ensemble systems. The purpose of this step is to understand where the systems are not able to learn the correct production rules, in order to improve them in the future.

Table 13
Reconstructing Latin proto-words. The first column indicates the modern language (or ensemble) that we used for training. For ensembles we report the results for the best performing ensemble. We report the average edit distance between the produced form and the correct form of the proto-word (EDIT) un-normalized (and in parentheses the normalized version), the coverage (COV for n ∈ {1, 5, 10}), and the mean reciprocal rank (MRR).

Language	EDIT	COV_1	COV_5	COV_{10}	MRR
Italian	2.45 (0.29)	0.14	0.32	0.35	0.22
Spanish	2.51 (0.29)	0.15	0.21	0.23	0.18
Portuguese	2.61 (0.30)	0.16	0.24	0.31	0.21
Ensemble	**2.31 (0.27)**	**0.22**	**0.32**	**0.42**	**0.27**
Oracle	1.76 (0.20)	0.28	0.41	0.47	0.34

(a) Data set 1 orthographic (Bouchard-Côté et al. 2007)

Language	EDIT	COV_1	COV_5	COV_{10}	MRR
Italian	2.52 (0.29)	0.16	0.28	0.32	0.21
Spanish	2.61 (0.30)	0.12	0.22	0.25	0.17
Portuguese	2.95 (0.34)	0.07	0.17	0.22	0.13
Ensemble	**2.28 (0.26)**	**0.14**	**0.32**	**0.36**	**0.21**
Oracle	1.93 (0.22)	0.23	0.36	0.39	0.30

(b) Data set 1 phonetic (Bouchard-Côté et al. 2007)

Language	EDIT	COV_1	COV_5	COV_{10}	MRR
Italian	1.57 (0.24)	0.25	**0.52**	**0.55**	0.37
Spanish	1.78 (0.27)	0.22	0.35	0.39	0.28
Portuguese	1.76 (0.28)	0.19	0.34	0.39	0.26
Romanian	2.12 (0.32)	0.18	0.31	0.36	0.25
French	2.31 (0.35)	0.13	0.24	0.30	0.18
Ensemble	**1.55 (0.23)**	**0.29**	0.49	**0.55**	**0.38**
Oracle	0.95 (0.14)	0.43	0.60	0.66	0.51

(c) Data set 2 (Reinheimer Ripeanu 2001)

Language	EDIT	COV_1	COV_5	COV_{10}	MRR
Italian	1.12 (0.14)	0.46	0.62	0.66	0.54
Spanish	1.31 (0.16)	0.42	0.59	0.61	0.49
Portuguese	1.30 (0.16)	0.41	0.58	0.61	0.49
Romanian	1.36 (0.16)	0.43	0.61	0.64	0.51
French	1.52 (0.18)	0.43	0.57	0.61	0.50
Ensemble	**1.07 (0.13)**	**0.50**	**0.70**	**0.74**	**0.59**
Oracle	0.65 (0.08)	0.66	0.77	0.79	0.71

(d) Data set 3 (see Section 4)

Looking at the incorrect productions that have one character different from the correct proto-word, we notice that sometimes the final consonant is mistaken (5% of the errors). Most commonly, *um* instead of *us* (4.1% of the errors): *serenum* instead of *serenus*, *cantum* instead of *cantus*, *novum* instead of *novus*. Another one-character mistake is sometimes failing to double a consonant (4% of the errors). For example,

Table 14
Example of reconstructing proto-words for the Latin proto-word *vicinus* (meaning *neighbor*). The correct productions are highlighted in **bold**.

Language	Word	5-best productions
French	voisin	vosinum, vosnum, vosine, vosinus, voiinum
Italian	vicino	vicinum, **vicinus**, vicenum, vicenus, vicnum
Portuguese	vizinho	vizinus, vizinum, **vicinus**, vizinium, vizinhum
Spanish	vecino	vecinum, vecinus, vicinum, vecenum, **vicinus**
Romanian	vecin	vicenus, vicenum, **vicinus**, vicinum, vecenus
Ensemble	all words	**vicinus**, vicinum, vicenus, vicenum, vecinus

ll or *ss*: *colapsus* instead of *collapsus*, *intervalum* instead of *intervallum*, *disociatio* instead of *dissociatio*, *esentia* instead of *essentia*.

For productions that have two characters different from the correct proto-word, we notice the following patterns in the incorrect productions: Sometimes the character *f* is mistakenly obtained instead of *ph*: *asfaltus* instead of *asphaltus*, *eufonia* instead of *euphonia*, *diafragma* instead of *diaphragma*. Another interesting pattern is obtaining the suffix *a* instead of *us*: *citrina* instead of *citrinus*, *alba* instead of *albus*. When this occurs for adjectives, the productions are not incorrect words in Latin; we obtain the feminine form instead of the masculine.

Comparison with Previous Work. Data set 1 allows us a fair comparison with the state-of-the-art method proposed by Bouchard-Côté et al. (2007), as illustrated in Table 16. On the same data set, the authors report 2.34 average edit distance between the produced words and the gold standard, when recreating Latin proto-words using the phonetic transcriptions. On the same data set, we obtain better results on both the orthographic and the phonetic version of the data set. The best of our systems (ensemble with fusion method based on the rank in the *n*-best lists and the training accuracy) obtains 2.28 average edit distance on the phonetic version of the data set and 2.31 average edit distance on the orthographic version. The oracle obtains 1.76 average edit distance on the orthographic version of the data set and 1.93 on the phonetic version.

The improvement of our systems is noteworthy because we are able to obtain these results with less data than in previous experiments—which in historical linguistics is essential, as resources are most of the time scarce: The system is able to perform well even when not having the phonetic transcripts of the words.

Additional Experiments: Neural Production of Related Words. We performed additional experiments using a recurrent neural network (RNN) system with an encoder–decoder

Table 15
Cognate sets with artificially reconstructed Latin proto-words.

Romanian	French	Italian	Spanish	Portuguese	Latin
uita	oublier	obliare	olvidar	olvidar	oblitare (forget)
forţă	force	forza	fuerza	força	fortia (force)
bară	bare	barra	barra	barra	barra (bar)
alb	blanc	bianco	blanco	branco	blancus (white)

Table 16
Comparison between previous work and our ensembles and oracles on Data set 1 (the lower the average edit distance between the correct proto-word and the production, the better).

System	Average edit distance
Bouchard-Côté et al. (2007)	2.34
Ensemble (orthographic)	2.31
Ensemble (phonetic)	2.28
Oracle (orthographic)	**1.76**
Oracle (phonetic)	1.93

architecture instead of a CRF. RNNs have been proven useful in many applications and are suitable for sequence labeling problems. However, they require large amounts of training data. In historical linguistics in general, and in the problem of reconstructing proto-words in particular, the resources are often scarce. Thus, we were interested to see whether RNNs can outperform the CRF system. We experimented with an encoder-decoder system with two long short-term memory layers, with stochastic gradient descent optimization and a global attention mechanism (Luong, Pham, and Manning 2015). We used the RNN implementation provided by TensorFlow (Luong, Brevdo, and Zhao 2017). We experimented with Word2Vec character embeddings as features, and also with embeddings extracted from the aligned words (similar to the features used for the CRF system).

The results of the RNN system are lower than those of the CRF system. The results are not included here, because they do not provide better performance. We leave for future work experimenting with additional features and set-ups, in order to compensate for the lack of training data.

6.3 Production of Modern Words

For borrowed words, we investigate the derivation of a word from a donor language into a recipient language (Dinu and Ciobanu 2017). Given the form of a word u in a donor language L_1, we develop a method that predicts the form v of the word u in a recipient language L_2, with the hypotheses that the word v will be derived in L_2 from the word u (through a borrowing process).

6.3.1 Experiments. We use the data set of related words introduced in Section 4, from which we extract Romanian words having etymons in 20 languages, covering all the European language families.[23] Romanian is a Romance language belonging to the Italic branch of the Indo-European language family. It is surrounded by Slavic languages and its relationship with the large Romance kernel was difficult. Its geographic position north of the Balkans put it in contact not only with the Balkan area but also with the vast majority of Slavic languages. Political and administrative relationships with the Ottoman Empire, Greece (the Phanariot domination), and the Habsburg Empire exposed Romanian to a wide variety of linguistic influences.

23 Romanian borrowed words from over 40 languages (Ciobanu and Dinu 2014a). In our experiments, we use the top 20 languages in terms of number of borrowed words, so that we have enough training data.

Table 17
Producing related words for borrowings, using lemmas as input. The recipient language is, in all cases, Romanian. We report the un-normalized average edit distance between the produced form and the correct form of the borrowing (and between parentheses the normalized version) and the coverage for the baseline and the proposed system.

Donor	Baseline				This Work			
Language	EDIT	COV_1	COV_5	COV_{10}	EDIT	COV_1	COV_5	COV_{10}
English	2.04 (0.23)	0.02	0.16	0.25	1.33 (0.15)	0.36	0.56	0.61
French	2.16 (0.24)	0.06	0.25	0.35	1.42 (0.15)	0.32	0.63	0.70
Italian	2.60 (0.32)	0.00	0.17	0.23	1.62 (0.23)	0.35	0.47	0.53
Latin	2.75 (0.34)	0.00	0.08	0.17	1.76 (0.22)	0.28	0.48	0.55
Neo-Greek	2.39 (0.29)	0.08	0.17	0.25	1.82 (0.24)	0.25	0.53	0.58
Old Slavic	2.34 (0.33)	0.08	0.18	0.23	1.84 (0.27)	0.17	0.39	0.47
German	2.36 (0.32)	0.07	0.23	0.26	2.00 (0.29)	0.26	0.41	0.45
Turkish	1.88 (0.27)	0.11	0.17	0.21	2.01 (0.29)	0.23	0.37	0.41
Bulgarian	2.33 (0.34)	0.06	0.20	0.21	2.22 (0.33)	0.15	0.23	0.28
Ruthenian	2.33 (0.35)	0.09	0.19	0.25	2.31 (0.35)	0.11	0.18	0.21
Russian	2.24 (0.33)	0.09	0.19	0.23	2.33 (0.33)	0.13	0.20	0.25
Albanian	2.60 (0.42)	0.06	0.11	0.12	2.35 (0.38)	0.08	0.20	0.25
Serbian	2.43 (0.37)	0.01	0.19	0.21	2.38 (0.36)	0.11	0.23	0.27
Polish	2.49 (0.38)	0.04	0.12	0.15	2.43 (0.36)	0.08	0.13	0.19
Portuguese	2.95 (0.52)	0.00	0.03	0.08	2.50 (0.43)	0.07	0.30	0.33
Slavic	2.88 (0.42)	0.05	0.11	0.17	2.66 (0.41)	0.12	0.27	0.31
Provençal	3.01 (0.49)	0.01	0.04	0.07	2.70 (0.44)	0.05	0.17	0.21
Hungarian	2.80 (0.43)	0.05	0.16	0.21	2.73 (0.42)	0.05	0.19	0.21
Spanish	3.22 (0.53)	0.02	0.06	0.11	3.06 (0.50)	0.05	0.12	0.15
Greek	4.36 (0.49)	0.01	0.08	0.08	4.28 (0.48)	0.05	0.15	0.15

We use a "majority class" type of baseline that does not take context into account. That is, it replaces each character in the source word with the most probably corresponding character in the target word, without taking previous and subsequent characters into account. The probabilities are extracted from the training data set.

6.3.2 Results and Analysis. First, we experiment using lemmas (dictionary word forms) as input. The results for this experiment are listed in Table 17. We observe that the best results are obtained from French and English donor words. The first eight languages that are ranked higher are those with which Romanian has the most intense cultural interaction, either more recently (English, for example), or in the past: in the period of the "re-Latinization" of Romanian (when the Italian and French influence was remarkable) or by continuous contact (with Turkish). The performance of producing word forms is lower even for related languages (such as Portuguese and Spanish); these languages are more remote from Romania, from a geographical point of view, and this might have made the contact between languages more difficult. What is more, for Spanish, for example, there has never been a significant Spanish cultural influence on Romanian.

The method we propose outperforms the baseline significantly for all considered languages except for Turkish, regarding the edit distance. Our assumption was that context is very relevant in producing related words and our hypothesis is confirmed by the difference in performance between our method and the baseline (since the baseline does not take context into account).

Table 18
Producing related words for borrowings in Romanian, using stems as input. For our method, we mark with * the results for which the difference to the experiment using lemmas as input is statistically significant (pairwise t-test, $p < 0.05$).

Donor	Baseline				This Work			
Language	EDIT	COV_1	COV_5	COV_{10}	EDIT	COV_1	COV_5	COV_{10}
French	1.65 (0.20)	0.27	0.44	0.49	1.43 (0.17)	0.41	0.56	0.61
English	1.94 (0.22)	0.18	0.34	0.39	1.48 (0.17)	0.38	0.51	0.56
Portuguese	2.13 (0.45)	0.06	0.20	0.23	1.67 (0.36)*	0.18	0.33	0.38
Italian	1.83 (0.27)	0.25	0.32	0.33	1.70 (0.26)	0.31	0.45	0.45
German	1.99 (0.29)	0.17	0.31	0.33	1.75 (0.27)*	0.28	0.49	0.51
Russian	2.18 (0.33)	0.12	0.20	0.29	1.97 (0.32)*	0.20	0.29	0.33
Turkish	1.96 (0.31)	0.13	0.23	0.25	2.12 (0.33)	0.20	0.30	0.33
Spanish	2.53 (0.46)	0.13	0.21	0.23	2.26 (0.44)*	0.15	0.23	0.26
Hungarian	2.48 (0.43)	0.08	0.16	0.21	2.38 (0.42)*	0.12	0.20	0.24

We further repeat this experiment with a modified version of the data set, in which we discard diacritics, in order to see if (and how) diacritics influence the learning process. The results are slightly improved when diacritics are not taken into account.

In a third experiment on producing modern words, we use, instead of lemmas, stems. Both lemmatization and stemming reduce the form of inflected or derived words to a common base word form, but stemming does this in a more drastic manner. That is, it reduces the words to much shorter forms than lemmatization. Furthermore, stemming might also remove prefixes from the words, which lemmatization does not. We believe that this might make a difference. We are interested to see if training and testing the system on stems, instead of lemmas, leads to better results. We use the Snowball Stemmer, which provides stemmers for 9 of our 20 donor languages. The results for this experiment are reported in Table 18. A possible explanation for the fact that stemming does not improve performance is that foreign influences, in the case of new words entering the language, can occur in the root of the words as well, and thus the root is not necessarily easier to produce than the entire word (that is, including affixes). This shows that Romanian is a complex language, partly because of the richness of its morphological changes. There have been many morpho-phonological changes across time, and in declinations and conjugations there are alternations even in the stem of the word (Radulescu Sala, Carabulea, and Contras 2015). Another possible explanation is that, by stemming, we lose information.

6.4 Production of Cognates

Further, we address the production of cognates. This task is very similar to producing related words for borrowings; the only difference is that instead of using a data set of *etymon-word* pairs, we use a data set of *cognate* pairs, extracted from the same data set of related words. In recent years, there has been a significant interest in identifying cognates using computational methods, but very few studies address the automatic production of the cognate pairs. Our purpose is to determine whether the system behaves differently, in terms of performance, for cognates compared to borrowings.

6.4.1 Experiments. In the first experiment, we identify five pairs of languages and their corresponding cognate pairs: Romanian–Spanish, Romanian–Italian, Romanian–Turkish,

Table 19
Producing related words for parallel lists of cognates. The first column indicates here the recipient language. The donor language is Latin in the first sub-table and French in the second one.

Language	Baseline				This Work			
	EDIT	COV_1	COV_5	COV_{10}	EDIT	COV_1	COV_5	COV_{10}
Italian	1.14 (0.14)	0.44	0.52	0.54	0.98 (0.12)	0.48	0.63	0.69
Romanian	2.37 (0.31)	0.03	0.22	0.40	1.06 (0.16)	0.47	0.61	0.66
Spanish	1.67 (0.21)	0.16	0.39	0.44	1.16 (0.15)	0.45	0.60	0.63
Portuguese	2.93 (0.33)	0.07	0.16	0.17	2.62 (0.30)	0.20	0.28	0.33

(a) Cognates with Latin ancestors

Language	Baseline				This Work			
	EDIT	COV_1	COV_5	COV_{10}	EDIT	COV_1	COV_5	COV_{10}
Romanian	1.90 (0.21)	0.14	0.26	0.28	0.86 (0.12)	0.46	0.73	0.77
Turkish	2.15 (0.27)	0.15	0.23	0.29	1.56 (0.19)	0.38	0.53	0.56

(b) Cognates with French ancestors

Romanian–Portuguese, and Romanian–English. We are interested in investigating if, for a given pair of languages, having one word from a cognate pair, we can automatically determine the orthographic form of its cognate pair.

Further, we take into account the common ancestor of the cognate pairs and investigate in which language the production is better. To this end, we extracted two data sets: one data set of Latin words that entered Romanian, Spanish, Portuguese, and Italian (Table 19a) and another data set of French words that entered Romanian and Turkish (Table 19b).

6.4.2 Results and Analysis. The results for the first experiment are reported in Table 20. For all five languages, the system performs, in both directions (i.e., from L_1 to L_2 and from L_2 to L_1) better than for deriving modern word forms from their ancestors. In Table 21 we report several examples of our system on producing cognates.

Table 20
Producing related words for cognates. For our method, we mark with * the results for which the differences to the first experiment are statistically significant (Mann-Whitney U-test, $p < 0.05$).

Lang.	Baseline				This Work			
	EDIT	COV_1	COV_5	COV_{10}	EDIT	COV_1	COV_5	COV_{10}
Es–Ro	1.71 (0.19)	0.10	0.34	0.41	0.91 (0.11)*	0.49	0.75	0.80
Ro–Es	1.60 (0.18)	0.12	0.19	0.22	0.94 (0.12)*	0.48	0.66	0.70
It–Ro	2.05 (0.23)	0.06	0.22	0.29	1.24 (0.15)	0.36	0.61	0.70
Ro–It	1.85 (0.21)	0.08	0.19	0.22	1.25 (0.14)*	0.43	0.56	0.64
Tr–Ro	1.30 (0.17)	0.27	0.32	0.35	1.28 (0.17)*	0.33	0.56	0.60
Ro–Tr	1.33 (0.18)	0.29	0.45	0.50	1.29 (0.17)*	0.41	0.61	0.65
Pt–Ro	1.54 (0.18)	0.17	0.42	0.50	1.29 (0.16)*	0.38	0.55	0.61
Ro–Pt	1.77 (0.21)	0.17	0.27	0.31	1.36 (0.16)*	0.32	0.51	0.55
En–Ro	2.01 (0.27)	0.01	0.14	0.19	1.30 (0.18)	0.35	0.62	0.70
Ro–En	2.08 (0.27)	0.03	0.13	0.18	1.47 (0.20)*	0.38	0.50	0.56

Table 21
Examples of producing cognates. The first column indicates the donor language. The recipient language is, in all cases, Romanian. The arrows (→, ←) indicate the direction of the production process. In the last column, we emphasize the true cognate (in **bold**).

Language	Cognate pair		Output (5-best list)
Italian	millenario - milenar (millenial)	→	**milenar**, milenarium, millenar, millenarium, milenariu
		←	milenario, milenare, milenarro, **millenario**, milanario
Spanish	petrificado - petrificat (petrified)	→	**petrificat**, petrificatum, petrificatus, petrificart, petrificant
		←	**petrificado**, petrificados, petrificacio, petrificación, petrificada
Portuguese	hipnose - hipnoză (hypnosis)	→	**hipnoză**, hipnosiune, ipnoză, ipnosiune, hipnos
		←	hipnoză, hipnosă, **hipnose**, hipnoser, hipnos
Turkish	otokrasi - autocraţie (autocracy)	→	autocrasie, **autocraţie**, otocrasie, autocracie, otocraţie
		←	otokrasyon, **otokrasi**, otokrasiyon, otokraţyon, otokrasyalamak

The results for the second experiment are reported in Table 19. We observe that for the first data set (Latin ancestors) the production is best for Italian, followed by Romanian and Spanish.

We perform the one-way ANOVA F-test ($p < 0.05$) with the null hypothesis $\mathcal{H}_0$: $\text{EDIT}_{Ro} = \text{EDIT}_{Es} = \text{EDIT}_{Pt} = \text{EDIT}_{It}$, where EDIT_L is the average edit distance between the produced and the correct word form, on the test set, for language L. Because the p-value is less than 0.05, we reject the null hypothesis. Further pairwise t-tests show that difference is statistically significant ($p < 0.05$) for Italian–Portuguese, Portuguese–Romanian, and Portuguese–Spanish. For the second data set (French ancestors), the system was able to learn orthographic patterns much better for Romanian than for Turkish. A pairwise t-test shows that the difference between the two languages is statistically significant ($p < 0.05$).

7. Conclusions

In this article, we designed tools to be used in historical linguistics, ran experiments on multiple data sets, and showed that they improve on the state-of-the-art results.

We described a dictionary-based approach to identifying cognates based on etymology and etymons. We accounted for relationships between languages and we extracted etymology-related information from electronic dictionaries.

We proposed a method for automatically identifying related words based on sequence alignment. We used aligned pairs of words to extract rules for lexical changes occurring when words enter new languages. We first applied our method for the task of identifying cognates on an automatically extracted data set of cognates for four pairs of languages. Then we applied the method for the task of discriminating between cognates and borrowings based on their orthography. Our results show that it is possible to identify the type of relationship with fairly good performance (over 85.0 accuracy for

three out of the four pairs of languages that we investigated). Our predictive analysis shows that the orthographic cues are different for cognates and borrowings, and that underlying linguistic factors captured by our model, such as affixes and diacritics, are indicative of the type of relationship between words. Other insights, such as the hyphenization or the part of speech of the words, are shown to have little or no predictive power. We intend to further account for finer-grained characteristics of the words and to extend our experiments to more languages. The method we proposed is language-independent, but we believe that incorporating language-specific knowledge might improve the system's performance.

We introduced an automatic method for the production of related words. We used an approach based on sequence labeling and sequence alignment, combining the results of individual systems using ensembles. We obtained fairly good results, and improved on previous work in this area. Our method has the advantage of requiring less input data than previous methods, and also accepting incomplete data, which is essential in historical linguistics, where resources are scarce. We first applied our method on multiple data sets of Romance languages, in order to reconstruct Latin proto-words. We conclude that leveraging information from multiple modern languages, in ensemble systems, improves the performance on this task, producing *n*-best list of proto-words to be further analyzed by linguists and to assist in the process of comparative reconstruction for endangered or extinct languages. Then we experimented with producing modern word forms for Romanian as a recipient language. We showed that languages are grouped, in the ranking, by their cultural influence on Romanian, rather than by the language families. We emphasize the difference in behavior between learning and producing borrowings, given their etymons (ancestors), and learning and producing cognates. The direction of the production does not seem to influence the results. Even when the output sequence does not match the true cognate, it might be a valid word in the recipient language. Sometimes, the produced sequences represent older forms of the words used today or, for nouns, the feminine form of the word. We observe that learning patterns from cognates lead to much better results than learning patterns from borrowings.

As future work, we intend to refine the fusion methods for the ensemble classifiers—as the oracle results showed the high potential of the approach—and to evaluate our method on other data sets that cover more languages (e.g., the Swadesh lists or the Austronesian basic vocabulary database [Greenhill, Blust, and Gray 2008]). We also intend to investigate further ways of improving the performance of the CRF system and to enhance the RNN system even with little data available.

Acknowledgments

We are grateful to the anonymous reviewers for their useful comments and suggestions. We would also like to thank Andreea Calude, Alexandra Cornilescu, Anca Dinu, and Ioan Pânzaru for their helpful feedback on an earlier version of this article.

References

Alkire, Ti and Carol Rosen. 2010. *Romance Languages: A Historical Introduction*. Cambridge University Press.

Ammar, Waleed, Chris Dyer, and Noah A. Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proceedings of the 4th Named Entity Workshop*, pages 66–70, Jeju Island.

Ashby, Michael and John Maidment. 2005. *Introducing Phonetic Science*. Cambridge University Press.

Atkinson, Quentin D. 2013. The descent of words. *Proceedings of the National Academy of Sciences*, 110(11):4159–4160.

Atkinson, Quentin D., 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.

Barbu, Ana Maria. 2008. Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 1937–1941, Marrakech.

Beinborn, Lisa, Torsten Zesch, and Iryna Gurevych. 2013. Cognate production using character-based machine translation. In *Proceedings of the 6th International Joint Conference on Natural Language Processing, IJCNLP 2013*, pages 883–891, Nagoya.

Bhargava, Aditya and Grzegorz Kondrak. 2009. Multiple word alignment with profile hidden Markov models. In *Proceedings of NAACL-HLT 2009, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 43–48, Boulder, CO.

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

Bouchard-Côté, Alexandre, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of NAACL 2009*, pages 65–73, Boulder, CO.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Bouchard-Côté, Alexandre, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of EMNLP-CoNLL 2007*, pages 887–896, Prague.

Brew, Chris and David McKelvie. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55, Ankara, Turkey.

Brodsky, David. 2009. *Spanish Vocabulary: An Etymological Approach*. University of Texas Press.

Buckley, Chris, Mandar Mitra, Janet A. Walz, and Claire Cardie. 1997. Using clustering and SuperConcepts within SMART: TREC 6. In *Proceedings of the 6th Text Retrieval Conference, TREC 1997*, pages 107–124, Gaithersburg, MD.

Campbell, Lyle. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Chang, Chih Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chitoran, Ioana. 2011. The nature of historical change. In *The Oxford Handbook of Laboratory Phonology*. Oxford University Press, pages 311–321.

Church, Kenneth W. 1993. Char Align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, ACL 1993*, pages 1–8, Columbus, OH.

Ciobanu, Alina Maria and Liviu P. Dinu. 2014a. An etymological approach to cross-language orthographic similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1047–1058, Doha.

Ciobanu, Alina Maria and Liviu P. Dinu. 2014b. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Volume 2: Short Papers*, pages 99–105, Baltimore, MD.

Ciobanu, Alina Maria and Liviu P. Dinu. 2014c. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043, Reykjavik.

Ciobanu, Alina Maria and Liviu P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 2: Short Papers*, pages 431–437, Beijing.

Ciobanu, Alina Maria and Liviu P. Dinu. 2018. Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 1604–1614, Santa Fe, NM.

Covington, Michael A. 1998. Alignment of multiple languages for historical comparison. In *Proceedings of ACL 1998, Volume 1*, pages 275–279, Montreal.

Delmestri, Antonella and Nello Cristianini. 2010. String similarity measures and PAM-like matrices for cognate identification. *Bucharest Working Papers in Linguistics*, 12(2):71–82.

Delmestri, Antonella and Nello Cristianini. 2012. Linguistic phylogenetic inference by PAM-like matrices. *Journal of Quantitative Linguistics*, 19(2):95–120.
Dijkstra, Ton, Franc Grootjen, and Job Schepens. 2012. Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15:157–166.
Dimitrescu, Florica. 1997. *Dictionar de Cuvinte Recente*. Ed. Logos.
Dinu, Liviu P. and Alina Maria Ciobanu. 2017. Romanian word production: An orthographic approach based on sequence labeling. In *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Revised Selected Papers, Part I*, pages 591–603, Budapest.
Eastlack, Charles L. 1977. Iberochange: A program to simulate systematic sound change in Ibero-Romance. *Computers and the Humanities*, 11:81–88.
Fellbaum, Christiane, editor . 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
Ganesh, Surya, Sree Harsha, Prasad Pingali, and Vasudeva Verma. 2008. Statistical transliteration for cross language information retrieval using HMM alignment model and CRF. In *Proceedings of the 2nd Workshop on Cross Lingual Information Access, CLIA 2008*, pages 42–47, Hyderabad.
Gomes, Luís and José Gabriel Pereira Lopes. 2011. Measuring spelling similarity for cognate identification. In *Proceedings of the 15th Portugese Conference on Progress in Artificial Intelligence, EPIA 2011*, pages 624–633, Lisbon.
Gooskens, Charlotte, Wilbert Heeringa, and Karin Beijering. 2008. Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, 2(1–2):63–81.
Gotoh, Osamu. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708.
Gray, Russel and Quentin Atkinson. 2003. Language tree divergences support the Anatolian theory of Indo-European origin. *Nature*, 426:435–439.
Greenhill, Simon J., Robert Blust, and Russel D. Gray. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271–283.
Hall, David and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of ACL 2010*, pages 1030–1039, Uppsala.
Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.
Hall, Robert Anderson. 1960. *Linguistics and Your Language*. Doubleday New York.
Hartman, Steven Lee. 1981. A universal alphabet for experiments in comparative phonology. *Computers and the Humanities*, 15:75–82.
Heggarty, Paul. 2012. In Beyond lexicostatistics: How to get more out of "word list" comparisons. In Soren Wichmann and Anthony P. Grant, editors, *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*. Benjamins, pages 113–137.
Hewson, John. 1974. Comparative reconstruction on the computer. In *Proceedings of ICHL 1974*, pages 191–197, Edinburgh.
Inkpen, Diana, Oana Frunza, 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, Borovets.
Jäger, Gerhard. 2018. Computational historical linguistics. *CoRR*, abs/1805.08099.
Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004*, pages 388–395, Barcelona.
Koehn, Philipp and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 711–715, Austin, TX.
Kondrak, Grzegorz. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000*, pages 288–295, Seattle, WA.
Kondrak, Grzegorz. 2002. Algorithms for Language Reconstruction. Ph.D. thesis, University of Toronto.
Kondrak, Grzegorz. 2004. Combining evidence in cognate identification. In *Proceedings of the 17th Conference of the*

Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, pages 44–59, London.
Kondrak, Grzegorz, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL*, pages 46–48, Edmonton.
Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, pages 282–289, Williamstown, MA.
Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.
List, Johann Mattis. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*, pages 117–125, Avignon.
List, Johann Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.
Luong, Minh-Thang, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. `https://github.com/tensorflow/nmt`.
Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP 2015*, pages 1412–1421, Lisbon.
Mackay, Wesley and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with pair hidden Markov models. In *Proceedings of the 9th Conference on Computational Natural Language Learning, CONLL 2005*, pages 40–47, Ann Arbor, MI.
McCallum, Andrew Kachites. 2002. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.
McMahon, April, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh sublists and the benefits of borrowing: An Andean case study. *Transactions of the Philological Society*, 103(2):147–170.
Melamed, Dan. 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, Cambridge, MA.
Minett, James W. and William S.-Y. Wang. 2003. On detecting borrowing: Distance-based and character-based approaches. *Diachronica*, 20(2):289–331.
Minett, James W. and William S.-Y. Wang. 2005. Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society*, 103(2):121–146.
Mulloni, Andrea. 2007. Automatic prediction of cognate orthography using support vector machines. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop, ACL 2007*, pages 25–30, Prague.
Mulloni, Andrea and Viktor Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2387–2390, Genoa.
Navlea, Mirabela and Amalia Todirascu. 2011. Using Cognates in a French-Romanian lexical alignment system: A comparative study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2011*, pages 247–253, Hissar.
Needleman, Saul B. and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.
Ng, Ee Lee, Beatrice Chin, Alvin W. Yeo, and Bali Ranaivo-Malançon. 2010. Identification of closely-related indigenous languages: An orthographic approach. *Int. J. of Asian Lang. Proc.*, 20(2):43–62.
Oakes, Michael P. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7:233–243.
Pagel, Mark, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.
Rama, Taraka. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka.
Rama, Taraka 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, pages 171–200.
Reinheimer Ripeanu, Sanda. 2001. *Lingvistica Romanica: Lexic, Morfologie, Fonetica*. Bucuresti.
Rădulescu Sala, Marina, Elena Carabulea, and Eugenia Contraş. 2015. *Formarea Cuvintelor în Limba Română*, volume 4. Editura Academiei Române.
Schuler, Gregory D. 2002. Sequence alignment and database searching. In A. D. Baxevanis and B. F. F. Ouellette, editors, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 43. John Wiley & Sons, Inc., pages 187–214.
Shawe-Taylor, John and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.
Simard, Michel, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, Montreal.
Smith, Temple F. and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.
St Arnaud, Adam, David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528, Copenhagen.
Steiner, Lydia, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.
Swadesh, Morris. 1954. Perspectives and problems of Amerindian comparative linguistics. *Word*, 10(2–3):306–332.
Tsvetkov, Yulia, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proceedings of NAACL-HLT 2015*, pages 598–608, Denver, CO.
Zhang, Ying, Almut Silja Hildebrand, and Stephan Vogel. 2006. Distributed language modeling for *n*-best list re-ranking. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 216–223, Sydney.