A Dictionary-Based Approach for Evaluating Orthographic Methods in Cognates Identification

Alina Maria Ciobanu
Faculty of Mathematics
and Computer Science
University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro

Liviu P. Dinu
Faculty of Mathematics
and Computer Science
Center for Computational Linguistics
University of Bucharest
ldinu@fmi.unibuc.ro

Abstract

In this paper we propose a method for identifying cognates based on etymology and etymons. We employ this approach to evaluate the extent to which lexical similarity can be used for automatic detection of cognate pairs. We investigate some orthographic approaches widely used in this research area and some original metrics as well. We apply this procedure for Romanian and its most closely related languages, French and Italian, but our method is applicable to any languages.

1 Introduction and Related Work

Cognates are words in different languages having the same etymology and a common ancestor. The task of cognates identification is widely used in historical and comparative linguistics, in the study of languages relatedness (Chin et al, 2010), phylogenetic inference (Atkinson et al, 2005) and in identifying how and to what extent languages changed over time. Besides these research areas, in which the genetic relationships between words are extremely relevant, cognates have been successfully used in other fields, such as language acquisition, bilingual word recognition (Dijkstra et al, 2012), corpus linguistics (Simard et al, 1992), cross-lingual information retrieval (Buckley et al, 1998) and machine translation (Knight et al, 2003). In these domains, the term "cognates" is usually used with a somewhat different meaning, denoting words with high orthographic/phonetic and cross-lingual meaning similarity, the condition of common etymology being left aside. Kondrak (2001) makes the distinction between the different interpretations of the notion and Inkpen et al (2005) present the definition of "genetic cognates".

In this paper we focus on genetic relationships between words and we use the term "cognates" in a broader meaning, counting as cognates the word-etymon pairs as well. Our motivation is that these pairs of words also share a common ancestor, thus complying with the cognates' definition. For example, the Romanian word *campion* (meaning *champion*) has Italian etymology and the etymon *campione*, which has Latin etymology and the etymon *campione(m)*. Thus, the Romanian word *campion* and the Italian word *campione* are cognates, as they share a common Latin ancestor.

The paper is organized as follows: we introduce our approach to cognates identification in Section 2. We describe the corpus used for our research in Section 3. We present several orthographic approaches used for cognates identification in Section 4. We evaluate these metrics and analyse the results of our experiments in Section 5. Finally, we draw some conclusion regarding our research in Section 6.

2 Our Approach

We focus on the Romanian language and we investigate its cognate pairs with two other Romance languages, French and Italian. We believe this comparison is interesting for the following reason: the two related languages differ significantly with respect to their orthographic depth: the mapping rules between graphemes and phonemes are more complex for French, which has a deep orthography, than for Italian, which has a highly phonemic orthography.

We identify the etymologies and etymons of the Romanian words using *dexonline* [1] machine-readable dictionary, which is an aggregator for over 30 Romanian dictionaries. By parsing its definitions, we are able to automatically extract information regarding words' etymologies and etymons. The most frequently used pattern is shown below.

[1] http://dexonline.ro

Proceedings of Recent Advances in Natural Language Processing, pages 141–147,
Hissar, Bulgaria, 7-13 September 2013.

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

As an example, we provide below an excerpt from a *dexonline* entry which uses this pattern to specify the etymology of the Romanian word *capitol* (which means *chapter*). When more options are possible for explaining a word's etymology, *dexonline* provides multiple etymologies. We account for all the given alternatives, enabling our method to provide more accurate results. In our example, the word *capitol* has double etymology: Latin (with the etymon *capitulum*) and Italian (with the etymon *capitolo*).

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

After determining the etymologies of the Romanian words, we translate in French all words without French etymology and in Italian all words without Italian etymology using *Google Translate* [2]. We consider cognate candidates pairs formed of Romanian words and their translations. Using French[3] and Italian[4] dictionaries, we extract etymology-related information for French and Italian words. To identify cognates we compare, for each pair of candidates, the etymologies and the etymons. If they match, we identify the words as being cognates. Our solution for addressing cognates identification answers Swadesh's question, as cited in (Campbell, 2003): "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, i.e. words that are known to be related.

For example, for the Romanian word *victorie*, *dexonline* reports 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 the Italian dictionary we identify, for this word, Latin etymology and the etymon *victoria*. We compare etymologies and 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.

3 The Corpus

We apply our method on a high-quality Romanian corpus comprising of the transcription of the parliamentary debates held between 1996 and 2007 in the Romanian Parliament, recently proposed in (Grozea, 2012). The sessions deal with a wide variety of topics regarding the political, social and economic fields. In this paper we decided to run our experiments using words extracted from a large corpus of transcribed spoken language, in order to investigate the cognates that are most frequently used in Romanian. This dataset covers particular cases in the task of cognates identification, such as cognates between which the degree of orthographic similarity is low (for example the Romanian word *atotputernicie*, which means *almightiness*, and its French cognate pair *omnipotence*, both sharing the Latin etymon *omnipotentia*) and vice versa, non-cognates that resemble one another (for example the Romanian word *mănăstire*, meaning *monastery* and having the Old Slavic etymon *monastyrí*, and its Italian translation *monastero*, having the Latin etymon *monasteriu(m)*).

Many words have undergone transformations by the augmentation of language-specific diacritics when entering a new language. From an orthographic perspective, the resemblance of words is higher between words without diacritics than between words with diacritics. For example, the similarity seems lower for the Romanian word *amiciţie* (which means *friendship*) and its French cognate pair *amitié* than for their corresponding forms without diacritics, *amicitie* and *amitie*. For this reason, we investigate the performances of the orthographic approaches to the task of cognates identification using two versions of the corpus: with and without diacritics included.

For preprocessing this corpus, we removed words that are irrelevant for our investigation, such

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

[3] http://www.cnrtl.fr/etymologie

[4] http://www.sapere.it/sapere/dizionari.html

as dates and numbers and all the transcribers' descriptions of the parliamentary sessions (such as *"The session began at 8:40."*), as we focus on the spoken language. We performed word segmentation, using whitespace and punctuation marks as delimiters, we lower-cased all words and we removed stop words, using a list of Romanian stop words provided by *Apache Lucene* [5] text search engine library . We lemmatized the words using *dexonline*, which provides information regarding the words' inflected forms and enables us to correctly identify lemmas where no part-of-speech or semantic ambiguities arise (in this case we consider the first occurred lemma).

4 Orthographic Approaches

Various word distances have been used in the task of string similarity computation. They have been applied in many different research areas, besides cognates identification, such as sentence alignment (Brew and McKelvie, 1996), record linkage (Jaro, 1989), stemming (Dalbelo and Snajder, 2009) and bioinformatics (Dinu and Sgarro, 2006). In (Kondrak, 2001) some of the most widely used measures are analysed, and their flaws and the differences between them are emphasized.

The approaches used to evaluate cognate pairs are divided in two groups: phonetic and orthographic. The orthographic approaches are usually used in corpus linguistics (Kondrak, 2001). We employ our method of identifying cognates to evaluate the extent to which lexical similarity can be used for automatic detection of cognates. We investigate some orthographic approaches widely used in this research area and some original metrics as well.

In (Inkpen et al, 2005) several orthographic similarity measures are used for the classification of pairs of words as cognates or false friends. For our investigation we chose some of the distances used in this paper, another distance that was successfully employed for record linkage and also an original metric in the field of cognates identification, rank distance.

- Levenshtein distance (Levenshtein, 1965), also named the edit distance, counts the minimum number of operations (insertion, deletion and substitution) required to transform one string into another. We use a normalized Levenshtein distance computed as:

$$EDIT(w_i, w_j) = \frac{LD(w_i, w_j)}{max(|w_i|, |w_j|)}$$

where $LD(w_i, w_j)$ is the Levenshtein distance for words w_i and w_j.

E.g. $\Delta(langue, lingua) = \frac{2}{6} = 0.33$

- Rank distance (Dinu and Dinu, 2005) is used to measure the similarity between two rank lists. A ranking of a set of n objects can be represented as a permutation of the integers $1, 2, ..., n$. S is a set of ranking results. $\sigma \in S$. $\sigma(i)$ represents the rank of object i in the ranking result σ. The rank distance is computed as:

$$RD(\sigma, \tau) = \sum_{i=1}^{i=n} |\sigma(i) - \tau(i)|$$

The ranks of the elements are given from bottom up, i.e. from n to 1, in a Borda order. The elements which do not occur in one of the rankings receive the rank 0. To extend the rank distance to strings, we index each occurence of a given letter a with a_k, where k is the number of its previous occurences, and then compute the rank distance for the new indexed strings which become in this situation rankings. In order to normalize it, we divide the obtained value by the maximum possible distance between two strings u and v, which is:

$$\frac{|u|(|u|+1)}{2} + \frac{|v|(|v|+1)}{2}$$

E.g. $\Delta(langue, lingua) = \frac{10}{42} = 0.23$

- Longest common subsequence ratio (Melamed, 1995) computes the similarity between two words dividing the length of the longest common subsequence of the two words by the length of the longer word:

$$LCSR(w_i, w_j) = \frac{LCS(w_i, w_j)}{max(|w_i|, |w_j|)}$$

where $LCS(w_i, w_j)$ is the longest common subsequence of w_i and w_j. We subtract this value from 1, in order to obtain the distance between two words.

E.g. $\Delta(langue, lingua) = 1 - \frac{4}{6} = 0.33$

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

- XDice (Brew and McKelvie, 1996) is a version of Dice's coefficient (Adamson and Boreham, 1972) which counts the number of shared character bigrams between two words and divides it by the number of bigrams in both words, allowing also extended bigrams (formed by the first and third letter of trigrams):

  $$XDICE(w_i, w_j) = \frac{2 * |xbi(w_i) \cap xbi(w_j)|}{|xbi(w_i) + xbi(w_j)|}$$

  where $xbi(w)$ is a function which determines the multi-set of character bigrams and extended bigrams in w. As XDice computes similarity between words, we subtract its value from 1 to obtain distances.

  E.g. $\Delta(langue, lingua) = 1 - \frac{2*4}{18} = 0.55$

- Jaro distance (Jaro, 1989) and its version, Jaro-Winkler distance (Winkler, 1990), are measures which account for the number and position of common characters between words. These metrics are described in (Delmestri and Dinu, 2012). Given two strings $w_i = (w_{i1}, ..., w_{im})$ and $w_j = (w_{j_1}, ..., w_{j_n})$, the number of common characters for w_i and w_j is the number of charachters w_{ik} in w_i which satisfy the condition:

  $$\exists w_{j_l} \text{ in } w_j : w_{ik} = w_{j_l}, |k - l| \leq \frac{max(m, n)}{2} - 1$$

  Let c be the number of common characters in w_i and w_j and t the number of character transpositions (i.e. the number of common characters in w_i and w_j in different positions, divided by 2). Jaro distance is defined as follows:

  $$J(w_i, w_j) = \frac{1}{3} * \left(\frac{c}{m} + \frac{c}{n} + \frac{c - t}{c} \right)$$

  As both Jaro and Jaro-Winkler metrics are string similarity measures, we subtract these values from 1 to obtain distances between words.

  E.g. $\Delta(langue, lingua) = 1 - \frac{1}{3} * \left(\frac{4}{6} + \frac{4}{6} + \frac{4-0}{4}\right) = 0.22$

  Jaro-Winkler distance accounts also for the length l of the common preffix of w_i and w_j ($l \leq 4$) and considers a scaling factor $p = 0.1$.

  $$JW(w_i, w_j) = J(w_i, w_j) + p * l * (1 - J(w_i, w_j))$$

  where $J(w_i, w_j)$ is the Jaro distance for words w_i and w_j.

  E.g. $\Delta(langue, lingua) = 1 - (0.77 + 0.1 * 1 * (1 - 0.77)) = 0.20$

5 Evaluation and Results Analysis

In order to evaluate the performances of these orthographic approaches to the task of cognates identification, we apply the method presented in Section 2 for determining cognate pairs in Italian and French for each word in the preprocessed corpus. The statistics for this phase of our procedure are listed in Table 1.

	Nwords	**Ncognates**	
		French	**Italian**
Type	162,399	77,029	35,581
Token	22,469,290	15,858,140	10,895,298
Lemmas	40,065	17,929	6,768

Table 1: Statistics for the Romanian corpus: the total number of type words, token words and lemmas (in column 1) and the number of type words, token words and lemmas having an etymon or a cognate pair in French (column 2) or in Italian (column 3). It can be noticed that the sum of token words with cognate pairs or etymons in French and Italian is higher than the total number of token words after preprocessing the corpus, due to the fact that many of these words have cognate pairs or etymons in both languages

Further, we excerpt from the corpus, for each of the two languages, random samples of 5,000 words which have a cognate pair in the related language and 5,000 which do not have such matching pair. We match these latter words with their translations. Thus, we obtain a sample of 10,000 pairs of words for Romanian and Italian, 5,000 pairs of cognates and 5,000 pairs of non-cognates. We obtain a similar set for Romanian and French. For each dataset we also consider the version without diacritics. We compute the lexical distances for each pair of words, setting various thresholds

French

th	EDIT				LCSR				RD				JW				XDICE			
	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F
0.0	06.4	100.0	53.2	12.0	06.4	100.0	53.2	12.0	06.4	100.0	53.2	12.0	06.4	100.0	53.2	12.0	06.4	100.0	53.2	12.0
0.1	08.9	94.3	54.2	16.3	09.3	93.8	54.4	17.0	15.2	87.6	56.5	26.0	41.9	81.1	66.1	55.3	09.4	92.5	54.3	17.0
0.2	24.9	83.2	60.0	38.4	26.4	82.5	60.4	40.0	40.6	83.4	66.3	54.7	71.8	78.6	76.1	75.1	18.1	83.1	57.2	29.8
0.3	47.6	83.1	68.9	60.5	50.3	82.3	69.7	62.4	63.3	81.1	74.3	71.1	88.2	75.9	**80.1**	81.6	34.0	81.8	63.2	48.0
0.4	68.7	80.6	76.1	74.2	71.8	79.4	76.6	75.4	79.7	78.5	78.9	79.1	95.6	71.1	78.3	81.5	49.1	80.6	68.7	61.0
0.5	84.9	78.2	80.6	81.4	87.1	76.4	**80.1**	81.4	89.9	75.5	**80.3**	82.0	98.2	62.7	69.8	76.5	65.4	79.5	74.3	71.8
0.6	91.3	76.0	**81.3**	83.0	93.2	73.1	79.4	81.9	94.4	71.3	78.2	81.2	99.4	54.3	57.9	70.2	74.7	78.4	77.1	76.5
0.7	94.8	72.9	79.8	82.4	96.4	67.4	74.9	79.3	97.2	65.3	72.7	78.1	99.4	53.3	56.1	69.4	81.8	77.1	78.8	79.4
0.8	98.2	65.1	72.8	78.3	98.8	57.5	63.0	72.7	98.5	58.7	64.6	73.6	99.4	53.2	56.1	69.3	89.9	74.3	**79.4**	81.4
0.9	99.4	57.1	62.4	72.6	99.7	52.2	54.1	68.5	99.5	54.0	57.3	70.0	99.4	53.2	56.1	69.3	94.5	69.2	76.3	79.9
1.0	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7

Italian

th	EDIT				LCSR				RD				JW				XDICE			
	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F
0.0	03.8	100.0	51.9	07.2	03.8	100.0	51.9	07.2	03.8	100.0	51.9	07.2	03.8	100.0	51.9	07.2	03.8	100.0	51.9	07.2
0.1	08.5	71.3	52.5	15.3	08.6	70.0	52.5	15.4	15.7	72.7	54.9	25.9	58.3	70.8	67.1	64.0	15.4	72.4	54.8	25.4
0.2	35.7	70.6	60.4	47.4	36.3	69.1	60.0	47.6	40.8	68.9	61.2	51.2	80.5	67.8	71.1	73.6	33.4	72.9	60.5	45.8
0.3	60.3	70.6	67.6	65.0	61.9	69.7	67.5	65.6	64.1	68.0	67.0	66.0	91.5	66.4	**72.6**	77.0	47.8	70.6	64.0	57.0
0.4	76.0	68.5	70.6	72.1	77.7	67.6	70.2	72.3	79.6	66.8	70.0	72.6	96.7	63.5	70.5	76.7	61.1	69.2	66.9	64.9
0.5	88.5	67.4	**72.8**	76.5	90.1	66.1	**72.0**	76.3	88.5	65.1	**70.6**	75.0	99.4	58.2	64.0	73.4	72.6	67.7	69.0	70.1
0.6	93.1	66.0	72.6	77.3	94.6	64.0	70.7	76.4	94.2	63.0	69.5	75.5	99.8	52.5	54.7	68.8	80.0	66.9	70.2	72.9
0.7	96.5	64.4	71.6	77.3	97.7	61.0	67.7	75.1	98.0	59.7	66.0	74.2	99.8	51.8	53.4	68.2	85.8	65.9	**70.7**	74.5
0.8	99.1	59.4	65.7	74.3	99.7	54.4	58.1	70.4	99.3	55.5	59.8	71.2	99.8	51.7	53.3	68.1	92.6	64.4	70.6	76.0
0.9	99.8	54.5	58.2	70.5	99.9	51.3	52.6	67.8	99.7	52.3	54.4	68.6	99.8	51.7	53.3	68.1	96.5	61.5	68.0	75.1
1.0	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7

Table 2: Recall (R), precision (P), accuracy (A) and f-score (F) values (computed as percentages) for orthographic measures in the task of cognates identification when diacritics are accounted for

for identifying cognates. The lists of cognates and non-cognates and the values computed by the orthographic distances for all the words in the Romanian-French and Romanian-Italian datasets are available from the authors on request. We count the occurences of each possible outcome: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In order to analyse and compare the relevance of these metrics, we further use these results to compute the values for recall, precision, accuracy and f-score using the formulas shown below, as presented in (Manning et al, 2008).

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$f - score = 2 * \frac{precision * recall}{precision + recall}$$

The results of our research are listed in Table 2 for the corpus with diacritics and in Table 3 for the corpus without diacritics. We highlighted the maximum accuracy obtained by each metric for thresholds between 0 and 1. Between Jaro and Jaro-Winkler distances, we decided to use only the latter metric in our analysis, as they are similar to a certain extent and we noticed that Jaro-Winkler distance provides better results.

According to the outcome of our investigation, the edit distance identifies Romanian-French and Romanian-Italian cognates with the highest degree of accuracy, reaching its maximum for a threshold value of 0.5 (and 0.6 for French, when diacritics are accounted for), followed closely by Jaro-Winkler distance and the longest common subsequence ratio. An interesting situation can be observed for Jaro-Winkler distance, whose accuracy decreses dramatically starting with 0.5 threshold, especially when diacritics are not taken into consideration. As expected, for each orthographic method the accuracy increases, reaches a maximum and then decreases, due to the precision-recall tradeoff. However, it is interesting to observe the similarity for the longest common subsequence ratio, rank distance and edit distance with regard to their accuracy curves when diacritics are accounted for. XDice and Jaro-Winkler distances exhibit different behaviours, in that Jaro-Winkler reaches its maximum accuracy for a threshold value lower than the average, while XDice has maximum accuracy for a higher threshold value. This behaviour stands for both languages.

It can be noticed that the orthographic approaches we used obtain higher degrees of accuracy for French than for Italian, which implies the fact that the orthographic changes undergone in the process of adapting to the Romanian language are a better indicator of cognacy for words with

French

th	EDIT				LCSR				RD				JW				XDICE			
	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F
0.0	08.9	100.0	54.4	16.3	08.9	100.0	54.4	16.3	08.9	100.0	54.4	16.3	08.9	100.0	54.4	16.3	08.9	100.0	54.4	16.3
0.1	12.3	94.0	55.8	21.7	12.9	93.2	56.0	22.6	21.4	87.7	59.2	34.4	58.1	80.6	72.0	67.5	13.4	90.3	56.0	23.3
0.2	34.1	81.2	63.1	48.0	35.9	80.6	63.6	49.7	54.6	82.5	71.5	65.7	82.6	77.9	79.6	80.2	28.3	81.8	61.0	42.1
0.3	60.5	82.0	73.6	69.6	62.9	81.0	74.1	70.8	73.4	79.9	77.4	76.5	92.5	74.5	**80.4**	82.5	48.8	80.6	68.5	60.8
0.4	77.1	79.8	78.8	78.4	79.3	78.2	78.6	78.8	85.4	77.1	**80.0**	81.1	96.7	69.4	77.0	80.8	63.8	79.5	73.7	70.8
0.5	89.1	77.1	**81.4**	82.7	90.9	75.0	**80.3**	82.2	92.3	73.4	79.4	81.8	98.8	60.6	67.3	75.1	76.4	78.5	77.7	77.4
0.6	93.9	74.8	81.1	83.3	95.3	71.2	78.4	81.5	95.5	68.9	76.2	80.0	99.5	53.6	56.7	69.7	82.5	77.3	79.1	79.8
0.7	96.5	71.4	78.9	82.1	97.6	65.3	72.9	78.3	97.8	62.7	69.9	76.4	99.6	52.6	55.0	68.9	87.5	75.6	**79.6**	81.1
0.8	98.5	63.1	70.5	76.9	99.1	55.8	60.3	71.4	98.9	56.7	61.8	72.1	99.6	52.6	54.9	68.8	93.0	72.2	78.6	81.3
0.9	99.6	55.6	60.0	71.3	99.8	51.6	53.0	68.0	99.7	52.9	55.4	69.1	99.6	52.6	54.9	68.8	96.7	66.6	74.1	78.9
1.0	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7

Italian

th	EDIT				LCSR				RD				JW				XDICE			
	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F	R	P	A	F
0.0	06.7	100.0	53.4	12.6	06.7	100.0	53.4	12.6	06.7	100.0	53.4	12.6	06.7	100.0	53.4	12.6	06.7	100.0	53.4	12.6
0.1	12.2	77.0	54.3	21.0	12.3	75.7	54.2	21.2	17.5	73.8	55.7	28.3	63.8	70.9	68.8	67.1	19.1	74.4	56.2	30.4
0.2	41.4	70.9	62.2	52.3	42.3	69.5	61.9	52.6	43.5	68.6	61.8	53.2	84.9	68.0	72.5	75.5	38.6	72.8	62.1	50.5
0.3	64.6	70.3	68.6	67.3	66.3	69.4	68.6	67.9	66.8	67.9	67.6	67.4	94.0	66.2	**73.0**	77.7	52.6	70.6	65.3	60.2
0.4	80.1	68.9	72.0	74.1	82.0	67.8	71.5	74.2	82.9	66.7	70.8	74.0	97.7	62.7	69.8	76.4	65.9	69.4	68.4	67.6
0.5	91.8	67.5	**73.8**	77.8	93.3	66.1	**72.7**	77.4	91.3	64.9	**70.9**	75.8	99.6	57.1	62.3	72.6	76.9	68.1	70.4	72.2
0.6	95.4	65.7	72.9	77.8	96.7	63.4	70.5	76.6	95.9	62.2	68.8	75.5	99.9	52.0	53.9	68.4	84.1	67.2	71.6	74.7
0.7	97.8	63.7	71.0	77.1	98.6	59.8	66.2	74.5	98.5	58.5	64.3	73.4	99.9	51.4	52.6	67.8	90.0	66.0	**71.9**	76.2
0.8	99.4	58.1	63.9	73.4	99.7	53.3	56.2	69.5	99.3	54.2	57.7	70.2	99.9	51.3	52.6	67.8	95.1	63.9	70.7	76.4
0.9	99.9	53.6	56.7	69.7	99.9	50.8	51.6	67.4	99.8	51.7	53.4	68.1	99.9	51.3	52.6	67.8	97.7	60.4	66.8	74.6
1.0	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7	100.0	50.0	50.0	66.7

Table 3: Recall (R), precision (P), accuracy (A) and f-score (F) values (computed as percentages) for orthographic measures in the task of cognates identification when diacritics are not accounted for

French etymons or cognate pairs than for words with Italian etymons or cognate pairs. A possible explanation is that starting with the 19th century numerous words were imported from French. That period represents a stage in the Romanian's language evolution in which norms for the vocabulary of the literary language were defined, including patterns for adapting neologisms to Romanian, and probably many of the French words which entered the language in the 19th century are in this situation.

6 Conclusion and Future Work

In this paper we proposed a dictionary-based approach to identifying cognate pairs. We extracted etymology-related information from online dictionaries and we accounted for etymologies and etymons to detect cognates. We applied our method on a high-volume Romanian corpus and we focused on detecting cognate pairs between Romanian and its most closely related languages, Italian and French. We used this method to investigate to which extent the lexical similarity can be used for automatic detection of cognates, analysing the performances obtained by various orthographic approaches: edit distance, rank distance, longest common subsequence ratio, XDice distance and Jaro-Winkler distance. Our results show that the edit distance classifies pairs of words as cognates or non-cognates with the highest degree of accuracy, obtaining better results for French than for Italian, with some improvements when diacritics are not accounted for.

A possible application for cognates identification is native language detection (Popescu and Ionescu, 2013). We believe that accounting for genetic relationships between words could prove useful for this task. In our future work we intend to further investigate the performances of the orthographic approaches to the task of cognates identification by introducing an additional step of parameter tuning for the threshold value in our procedure. We plan to apply this method of identifying cognates on the entire *dexonline* dictionary. In this paper we focused on the cognates that are most frequently used in Romanian, but we believe that obtaining an almost exhaustive dataset of Romanian-French and Romanian-Italian cognate pairs would be an important achievement.

Acknowledgements

The authors thank the anonymous reviewers for their helpful and constructive comments. The contribution of the authors to this paper is equal. Research supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

References

George.W. Adamson and Jillian Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10: 253–260.

Quentin D. Atkinson, Russell D. Gray and Geoff K. Nicholls, David J. Welch. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103: 193–219.

Chris Brew and David McKelvie. 1998. Word-pair extraction for lexicography. *Proceedings of the 2nd International Conference on New Methods in Language Processing, Ankara, Turkey*, 45–55.

Chris Buckley, Claire Cardie, Mandar Mitra and Janet Walz 1998. Using Clustering and SuperConcepts within SMART: TREC 6. *Proceedings of the Sixth Text Retrieval Conference (TREC-6).*

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

Beatrice Chin, Bali Ranaivo-Malançon, Ee-Lee Ng and Alvin W. Yeo. 2010. Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *International Journal of Asian Language Processing*, 20(2): 43–62.

Bojana Dalbelo Bašic and Jan Šnajder. 2009. String distance-based stemming of the highly inflected Croatian language. *Proceedings of the International Conference RANLP-2009. Borovets, Bulgaria*, 411–415.

Antonella Delmestri and Liviu P. Dinu. 2012. An Assessment of String Similarity Methods for Cognate Identification *In Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO), Belgrade, Serbia, April, Editors: Ivan Obradović, Emmerich Kelih, Reinhard Köhler*, 16–19.

Ton Dijkstra, Franc Grootjen and Job Schepens. 2012. Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, 15: 157–166.

Anca Dinu and Liviu P. Dinu. 2005. On the Syllabic Similarities of Romance Languages. *A. Gelbukh (ed.): CICLing 2005, Lecture Notes in Computer Science*, 3406: 785–789.

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

Diana Inkpen, Oana Frunza and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English *RANLP-2005, Bulgaria*, 251–257.

Cristian Grozea. 2012. Experiments and Results with Diacritics Restoration in Romanian. *TSD 2012*, 199–206.

Matthew A. Jaro. 1989. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Society*, 84(406): 414-–20.

Kevin Knight, Grzegorz Kondrak and Daniel Marcu. 2003. Cognates Can Improve Statistical Translation Models. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Companion volume of the Proceedings of HLT-NAACL 2003*, 46–48.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. *NAACL '01 Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8.

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

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Dan Melamed. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. *In Proceedings of the Third Workshop on Very Large Corpora.*

Marius Popescu and Radu Tudor Ionescu. 2013. The Story of the Characters, the DNA, and the Native Language. *In Proceedings of the BEA-8 Workshop of NAACL.*

Michel Simard, 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.*

William E. Winkler. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, 354—359.