Building a Dataset of Multilingual Cognates for the Romanian Lexicon

Alina Maria Ciobanu, Liviu P. Dinu

Faculty of Mathematics and Computer Science, University of Bucharest
Center for Computational Linguistics, University of Bucharest
alina.ciobanu@my.fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Identifying cognates is an interesting task with applications in numerous research areas, such as historical and comparative linguistics, language acquisition, cross-lingual information retrieval, readability and machine translation. We propose a dictionary-based approach to identifying cognates based on etymology and etymons. We account for relationships between languages and we extract etymology-related information from electronic dictionaries. We employ the dataset of cognates that we obtain as a gold standard for evaluating to which extent orthographic methods can be used to detect cognate pairs. The question that arises is whether they are able to discriminate between cognates and non-cognates, given the orthographic changes undergone by foreign words when entering new languages. We investigate some orthographic approaches widely used in this research area and some original metrics as well. We run our experiments on the Romanian lexicon, but the method we propose is adaptable to any language, as far as resources are available.

Keywords: cognates, etymology, orthographic approaches, Romanian

1. Introduction

Cognates are words in different languages having the same etymology and a common ancestor. Investigating pairs of cognates is very useful 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. In other several research areas, such as language acquisition, bilingual word recognition (Dijkstra et al., 2012), corpus linguistics (Simard et al., 1992), cross-lingual information retrieval (Buckley et al., 1997) and machine translation (Kondrak et al., 2003), the condition of common etymology is usually not essential and cognates are regarded as words with high cross-lingual meaning and orthographic or phonetic similarity.

In order to detect cognates, at least one dictionary containing etymology-related information is required for each of the considered languages. Electronic dictionaries enable the automatic or semi-automatic exploitation of the linguistic knowledge they comprise in various research areas, such as computational linguistics, lexicology and lexicography, natural language processing, artificial intelligence and data mining. The most useful electronic dictionaries in computational linguistics are machine-readable dictionaries (MRDs), which can be easily queried to retrieve data. For Romanian, we use *dexonline* MRD, which is described in detail in Section 3. The ideal situation is to use MRDs for all languages, but we are restricted in our investigation by the available resources. For foreign languages, we employ on-line dictionaries, we identify patterns and we use regular expressions to extract etymology-related information.

In this paper, we focus on etymology to identify cognates for the Romanian lexicon and we use the term *cognates* in a broader meaning, accounting for 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* (*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. 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 which 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 19th 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 rest of the paper is organized as follows: in Section 2. we analyse related work in this area. In Section 3. we describe and evaluate our method for building a dataset of multilingual cognates for the Romanian lexicon. In Section 4 we present one of the many possible applications for the obtained dataset, namely its usage as gold standard for the evaluation of orthographic approaches for cognates identification. Finally, in Section 5. we draw the conclusions of our study and briefly describe our plans for extending the method.

2. Related Work

There are three important aspects widely investigated in the task of cognates identification: semantic, phonetic and orthographic similarity. They were employed both individually (Simard et al., 1992; Inkpen et al., 2005; Church, 1993) and combined (Kondrak, 2004; Steiner et al., 2011) in order to detect pairs of cognates across languages. For determining semantic similarity, external lexical resources, such as WordNet (Fellbaum, 1998), are required. 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 et al., 2005; Hall and Klein, 2010); Levenshtein distance (Levenshtein, 1965), XDice distance (Brew and McKelvie, 1996) and longest common subsequence ratio (Melamed, 1995) are among the most frequently used metrics in this field.

Algorithms for string alignment were successfully used for cognates identification 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 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.

For Romanian, cognates among Romance languages were mostly investigated, due to Romanian's belonging to the Romance family. Rîpeanu (2001) built a parallel list of about 1,000 cognates with Latin ancestors for Romance languages. Navlea and Todirascu (2011) extracted Romanian-French cognate pairs from a legal parallel corpus and Ciobanu and Dinu (2013) extracted Romanian-French and Romanian-Italian cognate pairs from a high-volume Romanian corpus of transcribed parliamentary debates. However, to our knowledge, no such lists of cognates were built for the entire Romanian lexicon.

3. Cognates Identification

In this section we present and evaluate the method we used for building a dataset of cognates for Romanian and five other languages: French, Italian, Spanish, Portuguese and Turkish.

3.1. Method

Considering a set of words in a given language *L*, to identify the cognate pairs between *L* and a related language *L'* we apply the strategy proposed by Ciobanu and Dinu (2013): first, we determine the etymologies of the given words. Then, we translate in *L'* all words without *L'* etymology. We consider cognate candidates the pairs formed of input words and their translations. Using electronic dictionaries, we extract etymology-related information for the translated words. To identify cognates we compare, for each pair of candidates, the etymologies and the etymons. If they match, we identify the words as being cognates. 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 by disregarding final letters and we identify pairs such as *apostrof-appostrofo* and *admira-ammirare* as being cognates.

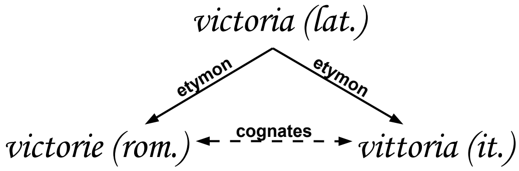

Figure 1: Example of cognates and word-etymon pairs

Our solution for addressing cognates identification answers Swadesh's question, as cited by 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. In Figure 1 we provide an example: for the Romanian word *victorie*, Romanian dictionaries report 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. Our method for cognates detection is represented in Algorithm 1 and Figure 2.

For determining words' etymologies we use *dexonline*[1] MRD, which is an aggregation of over 30 Romanian dictionaries. By parsing its definitions, we are able to automatically extract information regarding words' etymologies and etymons. The most frequently used pattern is shown below.

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

As an example, we provide below an excerpt from a *dexonline* entry which uses this pattern to specify the etymology of the Romanian word *exerciţiu* (*exercise*). For most words, etymological dictionaries offer a unique etymology, but when more alternatives are possible (there are words whose etymology was and remains difficult to ascertain), dictionaries may provide multiple etymological hypothesis. In our example, the word *exerciţiu* has double etymology: Latin (with the etymon *exercitium*) and French (with the etymon *exercice*).

```
<b>EXERCÍŢIU</b>
<abbr class="abbrev"
title="limba franceză">fr.</abbr>
<b>exercice</b>
```

[1] http://dexonline.ro

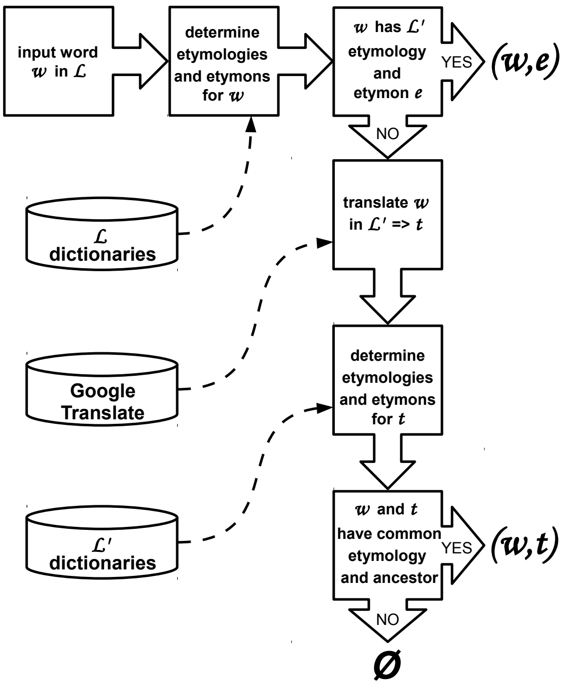
input word
w in $\mathcal{L}$
determine etymologies and etymons for w
w has $\mathcal{L}'$ etymology and etymon e
YES
(w,e)
NO
$\mathcal{L}$ dictionaries
translate w in $\mathcal{L}'$ => t
Google Translate
determine etymologies and etymons for t
$\mathcal{L}'$ dictionaries
w and t have common etymology and ancestor
YES
(w,t)
NO
Ø

Figure 2: Identification of word-etymon pairs and cognates for languages *L* and *L'*

Algorithm 1 Identification of word-etymon pairs and cognates for languages *L* and *L'*

```
Input: word $w$ in $L$, language $L'$
determine etymologies and etymons for $w$: $(L_{w_1}, e_{w_1}), ..., (L_{w_p}, e_{w_p})$
for each language $L_{w_k}$ in $T = \{L_{w_1}, ..., L_{w_p}\}$ do
    if $L_{w_k} = L$ then
        return $(w, e_{w_k})$
    end if
end for
translate word $w$ in $L'$; let $t$ be the translated word
determine etymologies and etymons for $t$: $(L_{t_1}, e_{t_1}), ..., (L_{t_q}, e_{t_q})$
for $(L_{w_i}, e_{w_i})$ in $P_w = \{(L_{w_1}, e_{w_1}), ..., (L_{w_p}, e_{w_p})\}$ do
    for $(L_{t_j}, e_{t_j})$ in $P_t = \{(L_{t_1}, e_{t_1}), ..., (L_{t_q}, e_{t_q})\}$ do
        if $L_{w_i} = L_{t_j}$ then
            if $e_{w_i} = e_{t_j}$ then
                return $(w, t)$
            end if
        end if
    end for
end for
return $\emptyset$
```

```
<abbr class="abbrev"
title="limba latină">lat.</abbr>
<b>exercitium</b>
```

For Italian[2], French[3], Spanish[4], Portuguese[5] and Turkish[6] we extract relevant etymology-related information from on-line dictionaries. We automatically simulate browser functionality and user actions using *HtmlUnit*[7] API, we parse HTML pages, we identify patterns and we use regular expressions to extract etymologies and etymons for foreign words. We manually translate Romanian words using *Google Translate*[8].

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 for each language. We obtain the following results: 95.45% accuracy for Romanian, 98% for Italian, for 96.6% for French, 98.2% for Spanish, 99,8% for Portuguese and, finally, 99.6% for Turkish.

3.2. Results

In Table 1 we report the number of words having an etymon or a cognate pair in each of the five considered languages. We account only for lexems, 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. 4,124 Romanian words in *dexonline* have an etymon or a cognate pair in all four Romance languages. The lists of cognates are available from the authors on request.

	♯words	♯etymons	♯cognates
FR	53,347	52,868	479
IT	13,377	9,874	3,503
ES	7,780	2,181	5,599
PT	10,972	1,318	9,654
TR	4,608	2,307	2,301

Table 1: Statistics for the Romanian lexicon

In Table 2 we provide statistics regarding the common ancestors of Romanian words and their cognates in French, Italian, Spanish, Portuguese and Turkish. In the 19th 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 (”pure” cognates) is much lower than the number of words with French etymons.

4. Application: Evaluation of Orthographic Approaches for Cognates Identification

We employ the dataset of cognates that we obtain as a gold standard for evaluating the performances of orthographic methods in the task of cognates identification. Detecting cognates based on etymology is useful and reliable, but for resource-poor languages more automated methods which require less linguistic knowledge might be necessary. We are interested in determining the extent to which lexical metrics can discriminate between cognates

[2] http://www.sapere.it/sapere/dizionari

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

[4] http://lema.rae.es/drae

[5] http://www.infopedia.pt/lingua-portuguesa

[6] http://www.nisanyansozluk.com

[7] http://htmlunit.sourceforge.net

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

	FR	IT	ES	PT	TR
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	-	-	-
Provencal	-	1	3	4	-
Russian	-	4	-	6	-
Spanish	-	34	-	72	-
Turkish	-	3	-	6	-
Total	479	3,503	5,599	9,654	2,301

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

and non-cognates, given the orthographic changes undergone by foreign words when entering new languages. Rules for adapting foreign words to the orthographic system of target languages might not have been very well defined in their period of early development, but they may have since become complex and specific. The orthographic approach relies on the idea that sound changes leave traces in the orthography and alphabetic character correspondences represent, to a fairly large extent, sound correspondences (Delmestri and Cristianini, 2010). Therefore, we believe these experiments are interesting and show one of the many possible applications for the dataset of cognates for the Romanian lexicon.

4.1. Similarity: Cognates vs. Etymons

A question that naturally arises is whether word-etymon pairs are closer, from an orthographic point of view, than cognate pairs. We compute the pairwise edit distance for related words and we report the overall results for cognates and word-etymon pairs in Table 3. For French, degrees of similarity are lower between cognate pairs than between word-etymon pairs, while for the other languages the opposite is true: words are closer to their etymons than to their cognate pairs. 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 lower between words with diacritics than between words without diacritics. For example, the similarity computed by subtracting the value of the normalized edit distance from 1 is lower for the Romanian word *amiciţie* (*friendship*) and its French cognate pair *amitié* than for their corresponding forms without diacritics, *amicitie* and *amitie*:

Sim(*amiciţie, amitié*) = 0.5 and Sim(*amicitie, amitie*) = 0.75,

For this reason, we report the similarity of word-etymon and cognate pairs in two versions of their forms: with and without diacritics.

	word-etymon pairs		cognate pairs	
	with diacritics	without diacritics	with diacritics	without diacritics
FR	0.72	0.77	0.62	0.69
IT	0.73	0.76	0.75	0.77
ES	0.53	0.57	0.76	0.79
PT	0.49	0.53	0.77	0.81
TR	0.63	0.69	0.74	0.76

Table 3: Word-cognate vs word-etymon overall pairwise similarity

4.2. Orthographic Approaches

We employ our method of identifying cognates 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 (edit distance, longest common subsequence ratio and XDice distance). Due to the limited space constraints, we only report the results obtained by edit distance, which achieved best performances overall.

We excerpt from the lexicon, for each of the five 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 each language, 5,000 pairs of cognates and 5,000 pairs of non-cognates. The only exception is Turkish, for which the number of cognates is lower than needed. Therefore, we select a random sample of 9,000 pairs of words, 4,500 pairs of cognates and 4,500 pairs of non-cognates for this language. We consider both versions of each dataset, with and without diacritics.

We split data into stratified train/test sets with a ratio of 4:1. We compute the normalized lexical distances for each pair of words. We follow a strategy similar to that proposed by Inkpen et al. (2005): we use the computed values as features, for each metric individually and we apply a fast decision tree learner implemented in *Weka*[9] workbench (Hall et al., 2009), namely *REPTree*. We set the value of the maximum tree depth to 1 and perform 10-fold cross-validation in order to select the best threshold for discriminating between cognates and non-cognates. Using the best threshold feature values selected for each metric and language, we further classify the pairs of words in our test set as cognates or non-cognates. In Table 4 we report the results obtained by edit distance on both train and test set.

Levenshtein distance obtains better performances than the other metrics, discriminating between cognates and non-cognates with highest accuracy in both versions of the test set, with and without diacritics. Highest accuracy values on both versions of the test set, with and without diacritics, are obtained for Turkish, which reaches a maximum of 89.5 accuracy, using the edit distance and a threshold of .6 when diacritics are not accounted for.

[9] http://www.cs.waikato.ac.nz/ml/weka

	FR	IT	ES	PT	TR
a)	78.1|.6	72.4|.6	73.6|.6	69.0|.6	87.7|.7
b)	77.5|.6	74.0|.5	73.6|.6	69.5|.6	88.5|.6
c)	76.6	71.1	74.4	69.5	88.1
d)	76.2	72.4	74.5	69.4	89.5

Table 4: Results for edit distance on train and test set:
a) *cross-validation accuracy|best threshold* when diacritics are accounted for;
b) *cross-validation accuracy|best threshold* when diacritics are not accounted for;
c) *accuracy* on test set when diacritics are accounted for;
d) *accuracy* on test set when diacritics are not accounted for

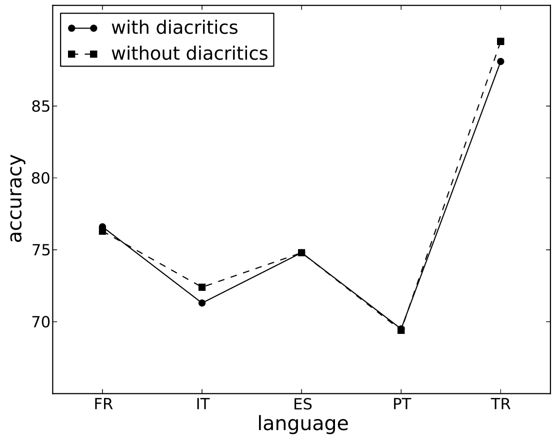

Figure 3: Highest accuracy for each language on both versions of the test set

5. Conclusion and Future Work

In this paper we proposed 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. As an application, we employed the dataset of cognates that we obtained as a gold standard for evaluating to which extent orthographic methods can be used to detect cognate pairs. Our results show that the edit distance discriminates between cognates and non-cognates with highest accuracy, obtaining slightly better performances when diacritics are not accounted for.

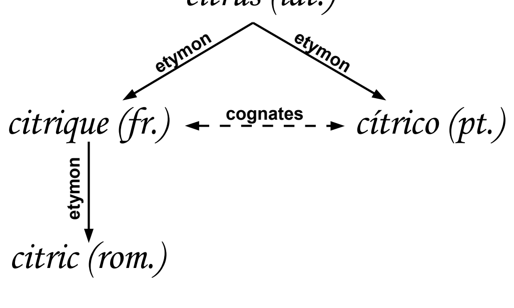

Figure 4: Example of more remotely related words

In our future work, we intend to develop a semi-automatic module for the word translation step in our method, based on a thorough preliminary analysis of the existing tools, such as GIZA++ (Och and Ney, 2003) or Moses (Koehn et al., 2007). We further plan to to extend our research to more languages, in order to cover a wider variety of linguistic families. Regarding our dataset of cognates, a possible improvement concerns the words which are more remotely related. In Figure 4 we provide an example: the Romanian word *citric* has French etymology and the etymon *citrique*, which has Latin etymology and the etymon *citrus*, and the Portuguese word *cítrico* has Latin etymology as well and the etymon *citru-*. Therefore, *citrique* and *cítrico* have a common etymon, while *citric* and *citrico* have a more remote common ancestor. In this paper we considered word-etymon pairs and cognates with common etymons, but words having common ancestors more remote in the line of descent are also worth being investigated.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. We thank Raluca Vasilache for the help with evaluating the cognates detection method. 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.

6. References

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

Chris Brew and David McKelvie. 1996. Word-Pair Extraction for Lexicography. In *Proceeding of the 2nd International Conference on New Methods in Language Processing, NeMLaP 1996*, pages 45–55.

Chris Buckley, 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.

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

Kenneth W. Church. 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.

Alina Maria Ciobanu and Liviu P. Dinu. 2013. A Dictionary-Based Approach for Evaluating Orthographic Methods in Cognates Identification. In *Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing, RANLP 2013*, pages 141–147.

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

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.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Osamu Gotoh. 1982. An Improved Algorithm for Matching Biological Sequences. *Journal of Molecular Biology*, 162(3):705–708.

David Hall and Dan Klein. 2010. Finding Cognate Groups Using Phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 1030–1039.

Mark Hall, 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.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions, ACL 2007*, pages 177–180.

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

Grzegorz Kondrak. 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.

Grzegorz Kondrak. 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.

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

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics Joint Workshop of LINGVIS and UNCLH*, pages 117–125.

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

Mirabela Navlea 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*, pages 247–253.

Saul B. Needleman 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.

Ee-Lee Ng, 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.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Sanda Reinheimer Rîpeanu. 2001. *Lingvistica Romanica: Lexic, Morfologie, Fonetica*. Ed. All. Bucuresti.

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.*

Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Lydia Steiner, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.