Automatic Detection of Cognates Using Orthographic Alignment

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

Words undergo various changes when entering new languages. Based on the assumption that these linguistic changes follow certain rules, we propose a method for automatically detecting pairs of cognates employing an orthographic alignment method which proved relevant for sequence alignment in computational biology. 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. Given a list of known cognates, our approach does not require any other linguistic information. However, it can be customized to integrate historical information regarding language evolution.

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 change 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.

The wide range of applications in which cognates prove useful attracted more and more attention on methods for detecting such related pairs of words. This task is most challenging for resource-poor languages, for which etymologically related information is not accessible. Therefore, the research (Inkpen et al., 2005; Mulloni and Pekar, 2006; Hauer and Kondrak, 2011) focused on automatic identification of cognate pairs, starting from lists of known cognates.

In this paper, we propose a method for automatically determining pairs of cognates across languages. 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. The rest of the paper is organized as follows: in Section 2 we present and analyze alternative methods and related work in this area. In Section 3 we introduce our approach for detection of cognates using orthographic alignment. In Section 4 we describe the experiments we conduct and we report and analyze the results, together with a comparison with previous methods. Finally, in Section 5 we draw the conclusions of our study and describe our plans for extending the method.

2 Related Work

There are three important aspects widely investigated in the task of cognate 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), or large corpora, 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 investi-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 99–105,
Baltimore, Maryland, USA, June 23-25 2014.

gated and compared (Inkpen et al., 2005; Hall and Klein, 2010); Levenshtein distance (Levenshtein, 1965), XDice (Brew and McKelvie, 1996) and the longest common subsequence ratio (Melamed, 1995) are among the most frequently used metrics in this field. Gomes and 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 employed in various approaches to cognate detection (Koehn and Knight, 2000; Mulloni and Pekar, 2006; Navlea and Todirascu, 2011). These orthographic changes have also been used in cognate production, which is closely related to the task of cognate detection, but has not yet been as intensively studied. While the purpose of cognate detection is to determine whether two given words form a cognate pair, the aim of cognate production is, given a word in a source language, to automatically produce its cognate pair in a target language. Beinborn et al. (2013) proposed a method for cognate production relying on statistical character-based machine translation, learning orthographic production patterns, and Mulloni (2007) introduced an algorithm for cognate production based on edit distance alignment and the identification of orthographic cues when words enter a new language.

3 Our Approach

Although there are multiple aspects that are relevant in the study of language relatedness, such as orthographic, phonetic, syntactic and semantic differences, in this paper we focus only on lexical evidence. 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).

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 which require less linguistic knowledge might be necessary. According to Gusfield (1997), an edit transcript (representing the conversion of one string to another) and an alignment are mathematically equivalent ways of describing relationships between strings. Therefore, because the edit distance was widely used in this research area and produced good results, we are encouraged to employ 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 cognate identification.

3.1 Orthographic Alignment

String alignment is closely related to the task of sequence alignment in computational biology. Therefore, to align pairs of words we employ the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970), which is mainly used for aligning sequences of proteins or nucleotides. Global sequence alignment aims at determining the best alignment over the entire length of the input sequences. The algorithm uses dynamic programming and, thus, guarantees to find the optimal alignment. 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).

3.2 Feature Extraction

Using 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 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 around mismatches. We experiment with two types of features:

i) n-grams around gaps, i.e., we account only for insertions and deletions;

ii) n-grams around any type of mismatch, i.e., we account for all three types of mismatches.

The second alternative leads to better performance, so we account for all mismatches. As for the length of the grams, we experiment with $n \in \{1, 2, 3\}$. We achieve slight improvements by combining n-grams, i.e., for a given n, we use all i-grams, where $i \in \{1, ..., n\}$. 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 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 8 features for the pair *(exhaustiv, esaustivo)*.

3.3 Learning Algorithms

We use Naive Bayes as a baseline and we experiment with Support Vector Machines (SVMs) to learn orthographic changes and to discriminate between pairs of cognates and non-cognates. 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 kernel (RBF), 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. Given two instances x_i and x_j, where $x_i \in \mathbb{R}^n$, the RBF kernel function for x_i and x_j is defined as follows:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0,$$

where γ is a kernel parameter.

We split the data in two subsets, for training and testing, with a 3:1 ratio, 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 $\{1, 2, ..., 10\}$ for c and over $\{10^{-5}, 10^{-4}, ..., 10^4, 10^5\}$ for γ. The values which optimize accuracy on the training set are reported, for each pair of languages, in Table 3.

4 Experiments

4.1 Data

We apply our method on an automatically extracted dataset of cognates for four pairs of languages: Romanian-French, Romanian-Italian, Romanian-Spanish and Romanian-Portuguese. In order to build the dataset, we apply the methodology proposed by Ciobanu and Dinu (2014) on the DexOnline[1] machine-readable dictionary for Romanian. 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. Finally, we obtain 445 pairs of cognates for Romanian-French[2], 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

[1] `http://dexonline.ro`

[2] The number of pairs of cognates is much lower for French than for the other languages because there are numerous Romanian words which have French etymology and, in this paper, we do not consider these words to be cognate candidates.

	1st	2nd	3rd	4th	5th
IT	iu>io	un>on	l->le	t$>-$	-$>e$
FR	un>on	ne>n-	iu>io	ţi>ti	e$>-$
ES	-$>o$	ţi>ci	–>ón	ie>ió	at>ad
PT	ie>ão	aţ>aç	ţi>çã	i$>-$	ă$>a$

Table 1: The most relevant orthographic cues for each pair of languages determined on the entire datasets using the χ^2 attribute evaluation method implemented in Weka.

	1st	2nd	3rd	4th	5th
IT	-$>e$	-$>o$	ă$>a$	–>re	ţi>zi
FR	e$>-$	un>on	ne>n-	iu>io	ţi>ti
ES	-$>o$	e$>-$	ţi>ci	ă$>a$	at>ad
PT	-$>o$	ă$>a$	e$>-$	-$>r$	-$>a$

Table 2: The most frequent orthographic cues for each pair of languages determined on the cognate lists using the raw frequencies.

comparison across languages, we keep 400 pairs of cognates and 400 pairs of non-cognates for each pair of languages. In Tables 1 and 2 we provide, for each pair of languages, the five most relevant 2-gram orthographic changes, determined using the χ^2 distribution implemented in Weka, and the five most frequent 2-gram orthographic changes in the cognate pairs from our dataset[3]. 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 1 and 2 is substitution.

4.2 Results Analysis

We propose a method for automatic detection of cognate pairs using orthographic alignment. We experiment with two machine-learning approaches: Naive Bayes and SVM. In Table 3 we report the results of our research. We report the n-gram values for which the best results are obtained and the hyperparameters for SVM, c and γ. The best results are obtained for French and Spanish, while the lowest accuracy is obtained for Portuguese. The SVM produces better results for all languages except Portuguese, where the accuracy is equal. For Portuguese, both Naive Bayes and SVM misclassify more non-cognates as cognates than viceversa. A possible explanation might be the occurrence, in the dataset, of more remotely related words, which are not labeled as cognates. We plan to investigate this assumption and to apply the proposed method on other datasets in our future work.

4.3 Comparison with Previous Methods

We investigate the performance of the method we propose in comparison to previous approaches for automatic detection of cognate pairs based on orthographic similarity. We employ 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)[4]. In addition, we use SpSim (Gomes and 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 dataset, we use the same train/test sets as we did in our previous experiments and we follow the strategy described in (Inkpen et al., 2005). First, we compute the pairwise distances between pairs of words for each orthographic metric individually, as a single feature[5]. 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, we further classify the pairs of words in our test sets as cognates or non-cognates. In Table 4 we report the results for each approach. Our method performs better than the orthographic metrics considered as individual features. Out of the four similarity metrics, SpSim obtains, overall, the best performance. These results support the relevance of accounting for orthographic cues in cognate identification.

[3]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.

[4]We use normalized similarity metrics. For the edit distance, we subtract the normalized value from 1 in order to obtain similarity.

[5]SpSim cannot be computed directly, as the other metrics, so we introduce an additional step in which we use 1/3 of the training set (only cognates are needed) to learn orthographic changes. In order to maintain a stratified dataset, 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 remaining of the initial training set for the next step of the procedure.

	Naive Bayes				SVM					
	P	R	A	n	P	R	A	n	c	γ
IT	0.72	0.93	79.0	1	0.76	0.92	81.5	1	1	0.10
FR	0.81	0.91	82.0	2	0.84	0.89	87.0	2	10	0.01
ES	0.79	0.92	84.0	1	0.85	0.88	86.5	2	4	0.01
PT	0.67	0.88	73.0	2	0.70	0.78	73.0	2	10	0.01

Table 3: Results for automatic detection of cognates using orthographic alignment. We report the precision (P), recall (R) 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.

	EDIT				LCSR				XDICE				SPSIM			
	P	R	A	*t*	P	R	A	*t*	P	R	A	*t*	P	R	A	*t*
IT	0.67	0.97	75.0	0.43	0.68	0.91	75.0	0.51	0.66	0.98	74.0	0.21	0.66	0.98	74.5	0.44
FR	0.76	0.93	82.0	0.30	0.76	0.90	81.5	0.42	0.77	0.79	78.0	0.26	0.86	0.83	85.0	0.59
ES	0.77	0.91	82.0	0.56	0.72	0.97	80.0	0.47	0.72	0.99	80.5	0.19	0.81	0.90	85.0	0.64
PT	0.62	0.99	69.5	0.34	0.59	0.99	65.5	0.34	0.57	0.99	63.5	0.10	0.62	0.97	69.0	0.39

Table 4: Comparison with previous methods for automatic detection of cognate pairs based on orthography. We report the precision (P), recall (R) and accuracy (A) obtained on the test sets and the optimal threshold *t* for discriminating between cognates and non-cognates.

5 Conclusions and Future Work

In this paper we proposed a method for automatic detection of cognates based on orthographic alignment. We employed the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) for sequence alignment widely-used in computational biology and we used aligned pairs of words to extract rules for lexical changes occurring when words enter new languages. We applied our method on an automatically extracted dataset of cognates for four pairs of languages.

As future work, we plan to extend our method on a few levels. In this paper we used a very simple substitution matrix for the alignment algorithm, but the method can be adapted to integrate historical information regarding language evolution. The substitution matrix for the alignment algorithm can be customized with language-specific information, in order to reflect the probability of a character to change into another. An important achievement in this direction belongs to Delmestri and Cristianini (2010), who introduced PAM-like matrices, linguistic-inspired substitution matrices which are based on information regarding orthographic changes. We plan to investigate the contribution of using this type of substitution matrices for our method.

We intend to investigate other approaches to string alignment, such as local alignment (Smith and Waterman, 1981), and other learning algorithms for discriminating between cognates and non-cognates. We plan to extend our analysis with more language-specific features, where linguistic knowledge is available. First, we intend to use the part of speech as an additional feature. We assume that some orthographic changes are dependent on the part of speech of the words. Secondly, we want to investigate whether accounting for the common ancestor language influences the results. We are interested to find out if the orthographic rules depend on the source language, or if they are rather specific to the target language. Finally, we plan to make a performance comparison on cognate pairs versus word-etymon pairs and to investigate false friends (Nakov et al., 2007).

We further intend to adapt our method for cognate detection to a closely related task, namely cognate production, i.e., given an input word *w*, a related language *L* and a set of learned rules for orthographic changes, to produce the cognate pair of *w* in *L*.

Acknowledgements

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

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*, 103:193–219.

Lisa Beinborn, 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.

Chris Brew and David McKelvie. 1996. Word-Pair Extraction for Lexicography. In *Proceeding of Text, Speech and Dialogue, TSD 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.

Chih-Chung Chang 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`.

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. 2014. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*.

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.

Luís Gomes 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. Software available at `http://research.variancia.com/spsim`.

Osamu Gotoh. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences: computer science and computational biology*. Cambridge University Press New York, NY, USA.

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. Software available at `http://www.cs.waikato.ac.nz/ml/weka`.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *5th International Joint Conference on Natural Language Processing, IJCNLP 2011*, pages 865–873.

Diana Inkpen, 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.

Philipp Koehn 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.

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 EACL 2012 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*.

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2387–2390.

Andrea Mulloni. 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.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2007. Cognate or False Friend? Ask the Web! In *Proceedings of the RANLP 2007 Workshop "Acquisition and Management of Multilingual Lexicons"*, pages 55–62.

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, RANLP 2011*, 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.

Gregory D. Schuler. 2002. Sequence Alignment and Database Searching. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 43. A. D. Baxevanis and B. F. F. Ouellette, John Wiley & Sons, Inc., New York, USA.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the 4th 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.