Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4096–4101
Marseille, 11–16 May 2020

Identifying Cognates in English-Dutch and French-Dutch by means of Orthographic Information and Cross-lingual Word Embeddings

Els Lefever, Sofie Labat and Pranaydeep Singh
LT3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{firstname.lastname}@ugent.be

Abstract

This paper investigates the validity of combining more traditional orthographic information with cross-lingual word embeddings to identify cognate pairs in English-Dutch and French-Dutch. In a first step, lists of potential cognate pairs in English-Dutch and French-Dutch are manually labelled. The resulting gold standard is used to train and evaluate a multi-layer perceptron that can distinguish cognates from non-cognates. Fifteen orthographic features capture string similarities between source and target words, while the cosine similarity between their word embeddings represents the semantic relation between these words. By adding domain-specific information to pretrained fastText embeddings, we are able to obtain good embeddings for words that did not yet have a pretrained embedding (e.g. Dutch compound nouns). These embeddings are then aligned in a cross-lingual vector space by exploiting their structural similarity (cf. adversarial learning). Our results indicate that although the classifier already achieves good results on the basis of orthographic information, the performance further improves by including semantic information in the form of cross-lingual word embeddings.

Keywords: cognate detection, multi-layer perceptron, orthographic similarity, cross-lingual word embeddings

1. Introduction

Cross-lingual similarity of word pairs from different languages concerns both formal and semantic overlap. Whereas the former refers to orthographic and/or phonetic similarity, the latter refers to translation equivalents in different languages (Schepens et al., 2013). Cognates are then defined as word pairs in different languages that have a similar form and meaning, which is often the result of a shared linguistic origin in some ancestor language (Frunza and Inkpen, 2007). Furthermore, true cognates can be distinguished from *false friends*, which have a similar form but different meaning, and from *partial cognates*, which share the same meaning for some, but not all contexts. For example, the English-Dutch word pair *father – vader* is a cognate pair, as both words in the pair have a similar form and meaning. In contrast, the French-Dutch *gras – gras* and the English-Dutch *driving – drijvende* are, respectively, instances of false friends and partial cognates. In the first word pair, the French *gras* means "fat, greasy", while the Dutch *gras* stands for "grass". In the second word pair, the English *driving* is defined as "communicating/ having great force" or "exerting pressure" by Merriam Webster[1], but it is also often used in contexts such as "driving a vehicule", while for Dutch, "drijvende" is not used in this latter context. Although cognates are often historically related, we chose to not incorporate this etymological criterion into the present study, as most previous research on computational cognate detection (see e.g. Schepens et al. (2013)) also disregards this criterion.

Cognate pair lists have shown to be useful for various research strands and applications. The use of cognates in second language learning has shown to accelerate the acquisition of vocabulary and to facilitate reading comprehension (Leblanc et al., 1989). Evidence from psycholinguistic experiments on vocabulary learning indeed points out that cognates are faster retrieved from memory and better remembered by learners (de Groot and Van Hell, 2005). Similarly, in translation tasks, cognates are translated faster and more correctly than other words (Jacobs et al., 2016). In Computer-Assisted Language Learning (CALL), tools have been developed to automatically annotate cognates and false friends in texts. Frunza and Inkpen (2007) implemented such a tool for French texts, in order to help second language learners of French (native English speakers).

In comparative linguistics, pairs of cognates can be employed to study language relatedness (Ng et al., 2010) or phylogenetic inference (Atkinson et al., 2010; Rama et al., 2018), whereas in translation studies, cognates and false friends contribute to the notorious problem of source language interference for translators (Mitkov et al., 2007). In NLP, finally, cognate information has been incorporated for various tasks, such as cross-lingual information retrieval (Makin et al., 2007), lexicon induction (Mann and Yarowsky, 2001; Sharoff, 2018) or machine translation (Kondrak et al., 2003; Jha et al., 2018).

The remainder of this paper is organized as follows. In Section 2., we present the existing approaches to cognate detection, which can be divided in three different strands: orthographic, phonetic and semantic methods. Section 3. gives a detailed overview of the created data set and the corresponding information sources, viz. orthographic and semantic similarity features, that are used for the classification experiments. In section 4., we report and analyze our experimental results, while Section 5. concludes this paper and gives some directions for future research.

2. Related Research

To build resources containing cognate information, manual work on cognate detection has been performed for var-

[1] https://www.merriam-webster.com/dictionary/driving

ious language pairs. As an example, we can refer to the work of Leblanc and Séguin (1996), who have collected 23,160 French-English cognate pairs from two general-purpose dictionaries (70,000 entries) and discovered that cognates make up over 30% of the French-English vocabulary. As the manual compilation of lists of cognate pairs is a very time-consuming and expensive task, researchers have started to develop automatic cognate detection systems. Three main approaches to cognate detection have been proposed, namely methods using (1) orthographic, (2) phonetic and (3) semantic similarity information. These information resources are either used individually or combined to perform automatic cognate detection (Mitkov et al., 2007; Kondrak, 2004; Schepens et al., 2013; Steiner et al., 2011).

Orthographic approaches view cognate detection as a string similarity task, and apply string similarity metrics such as the longest common subsequence ratio (Melamed, 1999) or the normalized Levenshtein distance (Levenshtein, 1965) to measure the orthographic resemblance between the candidate cognate pairs. **Phonetic** approaches also start from the idea of string similarity, but measure phonetic, instead of orthographic, similarity between cognate pairs. To this end, phonetic transcriptions of the words can be retrieved from lexical databases, and an adapted version of the standardised International Phonetic Alphabet (IPA) has been created to allow for cross-lingual comparison (Schepens et al., 2013).

Various algorithms were proposed for string alignment based on both the orthographic and phonetic form of the candidate cognates. Delmestri and Cristianini (2010) used basic sequence alignment algorithms, whereas Kondrak (2000) developed the ALINE system, which computes phonetic similarity scores using dynamic programming. List (2012), finally, proposed the LexStat framework, which combines different approaches to sequence comparison and alignment. More recently, machine learning approaches have been proposed for the task. Inkpen et al. (2005) mixed different measures of orthographic similarity using several machine learning classifiers, while Gomes et al. (2011) developed a new similarity metric able to learn spelling differences across languages. Ciobanu and Dino (2014) used aligned subsequences as features for machine learning algorithms to discriminate between cognates and non-cognates, whereas Rama (2016) explored the use of phonetic features to build convolutional networks to classify cognates.

Semantic similarity information has also been incorporated for the task of cognate detection (Mitkov et al., 2007). Taxonomies such as WordNet (Miller, 1995) have been used, starting from the intuition that semantic similarity between words can be approximated by their distance in the taxonomic structure. In addition, semantic similarity can also be computed by means of distributional information on the words. In this case, the intuition is that semantic similarity can be modelled via word co-occurrences in corpora, as words appearing in similar contexts tend to share similar meanings (Harris, 1954). Once the co-occurrence data is collected, the results are mapped to a vector for each word, and semantic similarity between words is then operationalized by measuring the distance (e.g. cosine distance) between their vectors.

In the proposed research, we build on the work of Labat and Lefever (2019) in which preliminary experiments were performed for English-Dutch cognate detection. Their pilot study showed promising results for a classifier combining orthographic similarity information with pretrained fastText word embeddings. In this research, we extend this work by (1) manually creating and annotating a gold standard for French-Dutch pairs of cognates, by (2) extending the word embeddings approach with domain- or corpus-specific information, and by (3) using more advanced methods to project the monolingual embeddings in a common cross-lingual vector space.

3. Cognate Detection System

We approached the task of cognate detection as a binary supervised classification task, which aims at classifying a candidate cognate pair as being COGNATE or NON-COGNATE. All the features described in the subsequent sections were treated as independent, and combined to train and test various classifiers for the task. Based on the results, we opted for a multi-layer perceptron (MLPClassifier) as implemented in the sklearn library in Python (Pedregosa et al., 2011). The classifier was evaluated with 5-fold cross-validation and a simple grid search was performed on the training folds to obtain optimal values for the hyperparameters. We found that the activation for a 3-layer network with 50, 100 and 150 neurons respectively, together with a constant learning rate of 0.001, worked optimally for the data at hand.

The rest of this section is structured as follows. Section 3.1. introduces the data set created to train a classifier for English-Dutch and French-Dutch cognate detection, while Section 3.2. describes in detail the two types of information sources used by the classifier, viz. orthographic and semantic similarity information between the source and target word of the candidate cognate pairs.

3.1. Data

To train and evaluate the cognate detection system, we created a context-independent gold standard by manually labelling English-Dutch and French-Dutch pairs of cognates, partial cognates and false friends in bilingual term lists. In this section, we describe how lists of candidate cognate pairs were compiled on the basis of the Dutch Parallel Corpus (Macken et al., 2011) and how a manual annotation was performed to create a gold standard for English-Dutch and French-Dutch cognate pairs.

To select a list of candidate cognate pairs, unsupervised statistical word alignment using GIZA++ (Och and Ney, 2003) was applied on the Dutch Parallel Corpus (DPC). This parallel corpus for Dutch, French and English consists of more than ten million words and is sentence-aligned. It contains five different text types and is balanced with respect to text type and translation direction. The automatic word alignment on the English-Dutch part of the DPC resulted in a list containing more than 500,000 translation equivalents. A first selection was performed by applying the Normalized Levenshtein Distance (NLD) (as implemented by Gries (2004)) on this list of translation equivalents and only considering equivalents with a distance smaller than or equal to 0.5. This resulted in a list with 28,503 English-Dutch candidate cognate pairs and 22,715 French-Dutch candidate cognate pairs, which were subsequently manually labeled. Our decision to apply the NLD threshold as a first filtering mechanism entails that word pairs are eliminated when they do not share the required orthographic similarity. This limitation of the current research was needed to make the manual annotation work practically feasible.

In order to create a gold standard for cognate detection, we applied the annotation guidelines that were established in Labat et al. (2019). The guidelines propose a clearly defined method for the manual labeling of the following six categories: (1) **Cognate**: words which have a similar form and meaning in all contexts, (2) **Partial cognate**: words which have a similar form, but only share the same meaning in some contexts, (3) **False friend**: words which have a similar form but a different meaning, (4) **Proper name**: proper nouns (e.g. persons, companies, cities, countries, etc.) and their derivations, (5) **Error**: word alignment errors and compound nouns of which one part is a cognate, but the other part is missing in one of the languages, and (6) **No standard**: words that do not occur in the dictionary of that particular language. The resulting gold standard for both language pairs is freely available for the research community (Labat, S. and Lefever, E., 2020)[2].

The data set used for the binary classification experiments consisted of COGNATE pairs (labels "cognate" and "partial cognate") and NON-COGNATE pairs (labels "error" and "false friend"). The categories of "proper name" and "no standard" were removed from the data set as they are almost always identical translations and would thus boost the performance of the system in an artificial way. Table 1 gives an overview of the distribution of the two classes in the gold standard data sets.

	Cognate	Non-cognate	Total pairs
GS English-Dutch	9,855	4,763	14,618
GS French-Dutch	8,146	2,593	10,739

Table 1: Distribution of the COGNATE and NON-COGNATE class labels in the two gold standards (GS) for English-Dutch and French-Dutch.

3.2. Information Sources

To train the binary cognate detection system, we combined orthographic and semantic similarity information in a multi-layer perceptron.

3.2.1. Orthographic Information

Fifteen different string similarity metrics were applied on the candidate cognates to measure the formal relatedness between source and target words. Eleven of these fifteen metrics were also used by Frunza et al. (2007). A detailed overview of all similarity metrics accompanied by a short definition is provided in Labat and Lefever (2019).

The following list summarizes the orthographic features implemented: (1) **Prefix** divides the length of the shared prefix by the length of the longest cognate in the pair, (2) **Dice** (Brew and McKelvie, 1996) divides the number of common bigrams times two by the total number of bigrams in the cognate pair, (3) **Dice (trigrams)** differs from Dice in that it uses trigrams instead of bigrams, (4) **XDice** is a variant of Dice as it uses bigrams that are created out of trigrams by deleting the middle letter in them, (5) **XXDice** incorporates the string positions of the bigrams into its metric, (6) **LCSR** stands for the longest common subsequence ratio, which is two times the length of the longest subsequence over the summed length of both sequences, (7) **NLS**, or the Normalized Levenshtein Similarity, equals one minus the minimum number of edits required to change one string sequence to another, (8-11) **LCSR (bigrams), NLS (bigrams), LCSR (trigrams), and NLS (trigrams)** differ from their standard metrics in that they use, respectively, bigrams and trigrams to calculate their results, (12) **Jaccard index** models the length of the intersection of both cognate strings over the length of the union of these strings, (13) **Jaro-Winkler similarity** is the complement of the Jaro-Winkler distance, (14-15) **Spsim option 1 and Spsim option 2** are the only metrics which require supervised training, in order to learn grapheme mappings between language pairs (Gomes and Pereira Lopes, 2011). They are trained by performing 5-fold cross-validation on the positive instances (i.e. cognates) in the data set.

3.2.2. Semantic Information

In addition to features modeling formal similarity between the source and target words, we also incorporated semantic information in our classifier. To this end, cross-lingual word embeddings were used, since these have been proven to work well for the cognate detection task in our pilot study on English-Dutch word pairs (Labat and Lefever, 2019). The former approach was improved in the following way.

Firstly, standard fastText word embeddings, which were pretrained on Common Crawl and Wikipedia and generated with the standard skip-gram model as proposed by Bojanowski et al. (2017), were extended with domain-specific word embeddings. This was accomplished by incrementally re-training the fastText embeddings with additional sentences from the Dutch Parallel Corpus to accommodate for new, unseen words (Grave et al., 2018). These words are mainly domain-specific and, consequently, absent from the Common Crawl and Wikipedia data. Incremental training of word embeddings is fairly common and has been explored in the past for a variety of models and domains (Kaji and Kobayashi, 2017).

Furthermore, in order to calculate similarities between words in the two different languages, the independently trained monolingual word embeddings have to be aligned in a common vector space. The development of cross-lingual mappings for monolingual word embeddings has been an

[2] http://semeval.lt3.ugent.be/cognates/

active research area in recent times. While initially, linear mapping method were proposed (see for instance Mikolov et al. (2013)), a lot of different ideas have been explored recently, such as minimization of Earth Mover's Distance (Zhang et al., 2017) or using the Wasserstein GAN as a means to minimize Sinkhorn distance (Xu et al., 2018). For our experiments, we used the approach proposed by Artetxe et al. (2018), because of its state-of-the-art results on downstream tasks such as word-for-word translation, which starts from the assumption that translations will have similar neighbors in the embedding space. This principle is used to define an initial parallel dictionary which is then iteratively corrected. The iterations involve a novel self-learning approach, which computes the optimal orthogonal mapping for the current dictionaries by means of Singular Value Decomposition (SVD). Subsequently, the dictionaries are improved with a modified version of the nearest neighbor algorithm.

The first bilingual dictionary is constructed by exploiting the almost identical spacial structure of cross-lingual synonyms in the embedding space. By iterative learning, the initially inducted bilingual dictionary can be extended after every iteration based on the current state of the alignment. After aligning the monolingual embeddings in a common vector space, the cosine similarity between the two words in question was calculated and used as an additional feature for classification.

It is worth noting that a number of words occur very sparsely in the corpus, as they have a frequency lower than 5. It is generally a good idea to not train embeddings for these words, since more context is required to not compromise the initial embeddings as well as the mappings. For English-Dutch, around 1,259 word pairs were ignored because of low frequencies, while for French-Dutch, around 1,482 word pairs were left out for similar reasons. Table 2 shows the class distribution of the data that was finally used for the experiments for English-Dutch and French-Dutch cognate detection.

	Cognate	Non-cognate	Total pairs
English-Dutch	8,886	4,473	13,359
French-Dutch	7,020	2,237	9,257

Table 2: Distribution of the COGNATE and NON-COGNATE class labels in data used for the English-Dutch and French-Dutch cognate detection experiments.

4. Results and Analysis

This section describes the classification performance for three different experimental setups: (1) a classifier incorporating fifteen orthographic similarity features, (2) a classifier incorporating a semantic similarity feature, which results from taking the cosine distance between the word embeddings of the words in the cognate pair, and (3) a classifier combining all orthographic similarity features with the semantic similarity feature. Table 3 lists the averaged precision, recall and F1-score for the three experiments performed for English-Dutch, whereas Table 4 lists the averaged precision, recall and F1-score for the same experiments on the French-Dutch data.

The experimental results reveal that for both English-Dutch and French-Dutch, the combined classifier incorporating orthographic and semantic similarity information outperforms the classifiers using only one type of information, viz. either orthographic or semantic information.

The results also show that the classifier only incorporating semantic information obtains very good results for the "COGNATE" class, whereas it obtains more moderate results, and especially low recall scores, for the "NON-COGNATE" class. As we only use one feature to capture semantic information, while we combine fifteen different orthographic features, the experimental results might be improved by adding additional semantic features and incorporating contextual word embeddings in future experiments.

A manual analysis of the output reveals some interesting cases that were misclassified by the learner that only uses orthographic string similarity metrics, but are correctly classified by the combined classifer. Examples of pairs that are now correctly classified as "NON-COGNATE" are *debut* (English) – *filmdebuut* (Dutch) and *inactive* (English) – *actieve* (Dutch), the former being a partial English compound and the latter being a pair of antonyms. Examples of pairs showing less orthographic similarity that are now correctly classified as "COGNATE" by adding semantic information are *leaves* (English) – *blaadjes* (Dutch), *self-regulation* (English) – *zelfregulering* (Dutch), *ankles* (English) – *enkels* (Dutch) and *weight* (English) – *gewicht* (Dutch).

5. Conclusion

This paper presents a novel gold standard and classification-based approach to binary cognate detection for English-Dutch and French-Dutch word pairs.

To distinguish cognates from non-cognates, a multi-layer perceptron is trained based on a combination of orthographic and semantic similarity features. To capture semantic similarity between the source and target words, monolingual word embeddings are created by adding domain-specific information to pretrained fastText embeddings. Subsequently, these monolingual embeddings are aligned in a cross-lingual vector space. Finally, the cosine distance between the source and target word is calculated and incorporated as a semantic similarity feature. The experimental results show that combining orthographic similarity features with cross-lingual word embedding information is a viable approach to cognate detection.

In future research, we plan to experiment with alternative word embedding methods and to perform trilingual machine learning experiments for cognate detection, combining the Dutch, French and English similarity information. This will enable us to gain insights into cross-lingual cognate detection. Additional experiments can also be performed using cross-lingual embeddings that are trained using a manually created bilingual dictionary to compare performance with the embeddings currently trained without any form of supervision. Finally, it would be interesting

	Cognates			Non-cognates			Average score		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Ortho	0.909	0.992	0.952	0.909	0.798	0.850	0.909	0.895	0.902
Sem	0.997	1.00	0.998	0.987	0.422	0.672	0.997	0.711	0.830
Ortho + Sem	0.915	0.993	0.955	0.915	0.793	0.853	0.915	0.893	0.904

Table 3: Precision (Prec), Recall (Rec) and F1-score for the classifiers incorporating the fifteen orthographic features (*Ortho*), the classifier incorporating only semantic information (*Sem*) and the classifier incorporating both orthographic and semantic similarity features (*Ortho + Sem*) for **English-Dutch**.

	Cognates			Non-cognates			Average score		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
Ortho	0.951	0.940	0.945	0.929	0.810	0.864	0.940	0.875	0.905
Sem	0.915	1.000	0.956	0.925	0.642	0.764	0.920	0.821	0.868
Ortho + Sem	0.943	1.000	0.971	0.943	0.804	0.879	0.943	0.908	0.925

Table 4: Precision (Prec), Recall (Rec) and F1-score for the classifiers incorporating the fifteen orthographic features (*Ortho*), the classifier incorporating only semantic information (*Sem*) and the classifier incorporating both orthographic and semantic similarity features (*Ortho + Sem*) for **French-Dutch**.

to perform multi-class experiments, where a distinction is made between cognates, false friends and non-related word pairs. To this end, a training and evaluation corpus containing cognate candidates in context will be built and manually annotated.

6. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Atkinson, Q., Gray, R., Nicholls, G., and Welch, D. (2010). From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103(2):193–21.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brew, C. and McKelvie, D. (1996). Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55.

Ciobanu, A. and Dinu, L. (2014). Automatic Detection of Cognates Using Orthographic Alignment. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference (Volume 2: Short Papers)*, pages 99–105.

de Groot, A. M. B. and Van Hell, J. (2005). The Learning of Foreign Language Vocabulary. In *Handbook of bilingualism: Psycholinguistic approaches*, pages 9–29. Oxford University Press.

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

Frunza, O. and Inkpen, D. (2007). A tool for detecting French-English cognates and false friends. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 91–100.

Gomes, L. and Pereira Lopes, J. G. (2011). Measuring Spelling Similarity for Cognate Identification. In L. Antunes et al., editors, *Progress in Artificial Intelligence*, pages 624–633. Springer.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487.

Gries, S. T. (2004). Shouldn't It Be Breakfunch? A Quantitative Analysis of Blend Structure in English. *Linguistics*, 42(3):639–667.

Harris, Z. S. (1954). Distributional Structure. *WORD*, 10(2-3):146–162.

Inkpen, D., Frunza, O., and Kondrak, G. (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.

Jacobs, A., Fricke, M., and F. Kroll, J. (2016). Cross-language Activation Begins During Speech Planning and Extends Into Second Language Speech. *Language Learning*, 66(2):324–353.

Jha, S., Sudhakar, A., and Kumar Singh, A. (2018). Neural Machine Translation based Word Transduction Mechanisms for Low-Resource Languages. *CoRR*, abs/1811.08816.

Kaji, N. and Kobayashi, H. (2017). Incremental Skip-gram Model with Negative Sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 363–371.

Kondrak, G., Marcu, D., and Knight, K. (2003). Cognates Can Improve Statistical Translation Models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Lin-*

guistics on Human Language Technology, HLT-NAACL 2003, pages 46–48.

Kondrak, G. (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*, pages 288–295.

Kondrak, G. (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.

Labat, S. and Lefever, E. (2019). A Classification-Based Approach to Cognate Detection Combining Orthographic and Semantic Similarity Information. In G. Angelova, et al., editors, *Proceedings of Recent Advances in Natural Language Processing (RANLP 2019)*, pages 603–611.

Labat, S., Vandevoorde, L., and Lefever, E. (2019). Annotation Guidelines for Labeling English-Dutch Cognate Pairs, version 1.0. Technical report, Ghent University, LT3 15-01.

Leblanc, R. and Séguin, H. (1996). Les congénères homographes et parographes anglais-français. In R. Courchêne, et al., editors, *Twenty-Five Years of Second Language Teaching at the University of Ottawa*, pages 69–91. Les presses de l'Université d'Ottawa.

Leblanc, R., Compain, J., Duquette, L., and Séguin, H. (1989). *L'enseignement des langues secondes aux adultes : recherches et pratiques*. Les presses de l'Université d'Ottawa.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

List, J.-M. (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.

Macken, L., De Clercq, O., and Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta*, 56(2):374–390.

Makin, R., Pandey, N., Pingali, P., and Varmain, V. (2007). Experiments in Cross lingual IR among Indian Languages. In *International Workshop on Cross Language Information Processing (CLIP-2007)*.

Mann, G. S. and Yarowsky, D. (2001). Multipath Translation Lexicon Induction via Bridge Languages. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 151–158.

Melamed, I. D. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, 25(1):107–130.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Mitkov, R., Pekar, V., Blagoev, D., and Mulloni, A. (2007). Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation*, 21(1):29–53.

Ng, E.-L., Chin, B., Yeo, A., and Ranaivo-Malançon, B. (2010). Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *Int. J. of Asian Lang. Proc.*, 20(2):43–62.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rama, T., List, J.-M., Wahle, J., and Jäger, G. (2018). Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2)*, pages 393–400.

Rama, T. (2016). Siamese convolutional networks based on phonetic features for cognate identification. *CoRR*, abs/1605.05172.

Schepens, J., Paterson, K., Dijkstra, T., Grootjen, F., and van Heuven, W. J. B. (2013). Cross-Language Distributions of High Frequency and Phonetically Similar Cognates. *PLOS ONE*, 8(5):1–15.

Sharoff, S. (2018). Language adaptation experiments via cross-lingual embeddings for related languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 844–849.

Steiner, L., Stadler, P., and Cysouw, M. (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127.

Xu, R., Yang, Y., Otani, N., and Wu, Y. (2018). Unsupervised Cross-lingual Transfer of Word Embedding Spaces. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945.

7. Language Resource References

Labat, S. and Lefever, E. (2020). *Gold Standard for Cognate Pairs in English-Dutch and French-Dutch*. LT3, Ghent University, 1.0, ISLRN 288-099-424-255-6.