REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2025 PUCV, Chile 58(118) 315-339

Detecting True Cognates and False Friends with Word Embeddings

Detección de cognados verdaderos y falsos amigos con word embeddings

Carlos Periñán-Pascual
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
ESPAÑA
jopepas3@upv.es

Nicolás José Fernández-Martínez
UNIVERSIDAD DE JAÉN
ESPAÑA
njfernan@ujaen.es

Received: 28-3-2024 / **Accepted:** 28-10-2024
DOI: 10.4151/S0718-09342025011801252

Abstract

Cognates are orthographically similar words in different languages that share the same etymology. Investigating cognates is valuable in fields such as historical linguistics, language acquisition, information retrieval, and machine translation, among many other. In this context, identifying false friends poses a challenge for automatic methods of cognate detection, as orthographic similarity is not sufficient to distinguish such word pairs. To this end, we evaluated seven different unsupervised vector-space models based on neural networks to detect cognates in general and to distinguish true cognates from false friends in a list of word pairs in English and Spanish. This variety of models allowed us to determine the impact of several factors on the quality of the results and the effectiveness of the models: language resources employed in model construction (e.g., text corpora, lexical associative networks, or both), cross-lingual alignment of semantic spaces, and meaning conflation in polysemous words.

Keywords: cognate, false friend, word embedding, semantic similarity

Resumen

Los cognados son palabras ortográficamente similares en distintas lenguas que comparten la misma etimología. La investigación de los cognados es valiosa en ámbitos como la lingüística histórica, la adquisición de lenguas, la recuperación de información y la traducción automática, entre otros. En este contexto, la identificación de *false friends* supone un reto para los métodos automáticos de detección de cognados, ya que la similitud ortográfica no es suficiente para detectar estos pares de palabras. Con este fin, evaluamos siete modelos vectoriales-espaciales no supervisados diferentes basados en redes neuronales para detectar cognados y distinguir los cognados verdaderos de los *false friends* a partir de una lista de pares de palabras en inglés y español. Esta variedad de modelos permitió determinar el impacto de varios factores en la calidad de los resultados y la eficacia de los modelos: los recursos lingüísticos empleados en la construcción del modelo (por ejemplo, corpus de textos, redes asociativas léxicas o ambos), la alineación interlingüística de los espacios semánticos y la fusión de significados en palabras polisémicas.

Palabras clave: cognado, *false friend*, vectores semánticos, similitud semántica

INTRODUCTION

True cognates and false friends have been used in various research fields and applications. For example, an extensive list of false friends can be useful not only for foreign-language learners (Procter, 1995), but also for linguists that study language relatedness (Ng et al., 2010). Integrating a list of false friends into computer-assisted language learning (Frunza & Inkpen, 2007) or proofreading software (Miłkowski, 2010) also has shown to be very useful. Moreover, the use of cognates accelerates the process of corpus alignment (Nazar, 2011), bilingual lexicon construction (Gurrutxaga et al., 2006) or machine translation (Kondrak et al., 2003) in natural language processing (NLP). In the last few years, research on true cognates and false friends continues to show an active interest in fields such as translation and second language acquisition (Hansen-Schirra et al., 2017; Otwinowska et al., 2020), where cognates and false friends can contribute to the problem of source language interference (i.e., negative transfer). However, as explained by Mitkov et al. (2007), obtaining a list of cognates and false friends is difficult, especially for poorly spoken languages.

The compilation of such lists is a time-consuming and labour-intensive task. Therefore, the solution to this lexicographical bottleneck can be found in automatically identifying a large number of true cognates and false friends from lexical data. In this context, this research proposes a method for automatically recognising true cognates and false friends for a given pair of languages. In particular, the main contribution of this article is to assess the impact of three aspects of neural-network models (i.e., language resource type, cross-lingual embedding alignment, and meaning conflation) in the semantic-similarity stage of this task. Although the experiment is focused on English and Spanish words, the proposed method can be used for a variety of languages.

The remainder of this article is organised as follows. Section 1 describes the most relevant works for this study. Then, section 2 provides an accurate account of the proposed research method. Subsequently, section 3 describes the experiment, whereas section 4 shows the results. After that, section 5 interprets the empirical evidence. Finally, we present our conclusions.

1. Definitions

This section explains the main linguistic terms related to our research topic to clarify theoretical confusion: (true, partial, false or deceptive) cognate and (chance and semantic) false friend. The term "cognate" originates from *cognatu(m)* in Latin, which results from *cum* [with] and *natus* [born], thus meaning "born together", i.e., having the same etymology. Cognates are etymologically related words across languages that "share (parts of) their orthographic and/or phonological form" (De Groot & Keijzer, 2000, p. 3), leading to a similar spelling and/or pronunciation (Dijkstra et al., 2010). Therefore, two or more formally similar words that have the same origin and the same meaning in two or more languages should be called "true cognates" (e.g., *colour* [English] and *color* [Spanish], both from Latin *color*) (De Groot, 2011). Considering historically related words, Granger (1993) made a distinction between "good cognates" (i.e., true cognates), which have the same meaning, and "deceptive cognates", which have partially or totally different meanings. Moreover, Sabino (2002) distinguished between "false cognates", i.e. which have different meanings and etyma (e.g. *ape* [*monkey* in English, from Old Saxon *ape*] and *ape* [*bee* in Italian, from Latin *apem*]), and "deceptive cognates", i.e., which have different meanings but the same etymon (e.g., *library* ["place from which books can be borrowed" in English] and *librairie* [*bookshop* in French], both from Latin *librarium*). Sabino (2016) also explained that the set theory could be used to represent these categories of cognates accurately. Suppose that A and B are two sets containing the meanings of two lexical items in two different languages. True cognates can be represented as equal sets (i.e., $A = B$), false cognates as disjoint sets (i.e., $A \cap B = \emptyset$), partial cognates as inclusion subsets (i.e., $A \subset B \,\&\, A \neq B$), and deceptive cognates as intersection sets (i.e., $A \cap B \neq \emptyset$). Finally, in case two words have a similar form but only share the same meaning in some contexts, Labat, Vandevoorde and Lefever (2019) considered them "partial cognates". For example, *argument* means either dispute or a set of reasons in English, but only the second meaning is expressed by *argumento* in Spanish, although both words come from the Latin word *argumentum*.

On the other hand, a commonly accepted term in translation and foreign-language learning literature is "false friend", an umbrella term for lexical items with similar spellings or pronunciations, but different meanings since they have different semantic histories in their corresponding languages. De Groot (2011) subcategorised false friends based on the type of form overlap (i.e., phonological or orthographic) and the degree of form overlap (i.e., complete or for the larger part). In this regard, false friends that completely overlap in phonology or orthography are called interlexical homophones and interlexical homographs, respectively. On the other hand, false friends with a large phonological or orthographic overlapping are called interlexical homophonic neighbours and interlexical homographic neighbours, respectively. Dominguez and Nerlich (2002) drew a distinction between "chance false friends" (i.e.

false cognates) and "semantic false friends" (i.e,. partial and deceptive cognates). In turn, semantic false friends can be divided into two groups: full false friends (i.e., deceptive cognates), where the meanings of two words have diverged widely (e.g,. *fastidious* [English] and *fastidioso* [Spanish]), and partial false friends (i.e. partial cognates), where some of the meanings of the words are different but others remain the same (e.g., *professor* [English] and *profesor* [Spanish]).

As in most NLP studies about this topic, our research does not consider the etymology of words. Therefore, we focus on two binary classification tasks: the first one, cognate detection, where orthographically similar words in different languages are found, and the second one, cognate classification, where the challenge lies in recognising whether such words are translation equivalents (i.e., true cognates) or not (i.e., partial, false or deceptive cognates: false friends).

2. Related work

Most of the work aimed at identifying true cognates and false friends is based on some model that determines the similarity between words in different languages. The main difference lies in the type of similarity (i.e., orthographic, phonetic, semantic, or a combination of them), and the type of resources that provide evidence to assess such similarities (e.g., corpora, dictionaries, and thesauri).

On the one hand, Schepens et al. (2012), who studied the distribution of cognates in six European languages, adopted the orthographic approach to cognate recognition. In particular, they performed two steps: extracting word pairs directly from a translation lexicon, and classifying the extracted pairs as cognates or non-cognates based on orthographic similarity. A popular technique to assess orthographic similarity is the Levenshtein Distance (Levenshtein, 1965), which counts the minimal number of substitutions, insertions, and deletions required to edit one string into another. The problem of this metric lies in the fact that it computes high scores for long words and low scores for short words, so it was modified to become the Normalised Levenshtein Distance, as found in Inkpen and Frunza (2005) or Schepens et al. (2012), among others. Another effective technique is the Longest Common Subsequence Ratio (LCSR) of two tokens, which is "the ratio of the length of their longest (not necessarily contiguous) common subsequence (LCS) and the length of the longer token" (Melamed, 1999, p. 113).

On the other hand, one of the most representative works of the phonetic approach is Kondrak (2000), who devised an algorithm for aligning phonetic sequences (ALINE). Indeed, it was used to detect cognates by assigning a similarity score to pairs of phonetically transcribed words, where each phoneme was represented as a vector of phonetically based feature values. Kondrak and Dorr (2004) enhanced ALINE with orthographic evidence to increase accuracy in the task of cognate recognition. Palmero Aprosio et al. (2020) explained that orthographic similarity proves effective for typologically similar languages, but some phonetic similarity technique is required when the words involved belong to languages that do not share the same writing system.

Finally, the semantic approach to cognate recognition usually determines similarity based on information from corpora and knowledge bases, apart from relying on orthographic and phonetic evidence. For example, distributional information can be obtained through the local context in which candidate cognates occur in corpora. Through this approach, a lexical item that appears in a specific context can have a true cognate if the other word in the pair occurs in a similar context, hence extensive corpora are required. Moreover, information from knowledge bases, e.g., WordNet (Fellbaum, 2010), serves to determine the semantic similarity between two candidate cognates by considering their presence in glosses (e.g., Kondrak, 2001) or their position in a hierarchical taxonomy (e.g., Mitkov et al., 2007; Mulloni et al., 2007). Some of the most representative works are described as follows.

Brew and McKelvie (1996) detected a large number of English–French word pairs based on their occurrence in similar local contexts in a multilingual, non-annotated corpus of parliamentary questions, and then identified cognates and false friends relying on orthographic similarities.

Kondrak (2001) combined ALINE with a procedure that considers WordNet-based semantic similarity between words by analysing the information extracted from glosses.

In Mulloni et al. (2007), cognate recognition started with a purely orthographic approach and then leveraged semantic evidence from thesauri and monolingual corpora. Orthographically similar word pairs were detected through transformation rules. Semantic similarity resulted from combining the taxonomic similarity computed from EuroWordNet (Vossen, 1998) with the distributional similarity from corpora.

Nakov et al. (2007) detected true cognates and false friends relying solely on semantic similarity, ignoring orthographic or phonetic evidence. In particular, they collected semantic information about the local context of candidates by using the Web as a corpus, rather than pre-existing corpora, and a bilingual glossary, where word translations serve as cross-linguistic "bridges". Contextual semantic vectors were compared to assess semantic similarity.

In Mitkov et al. (2007), candidate pairs were extracted through distributional evidence from non-parallel bilingual corpora, and then the extracted pairs were classified as true cognates or false friends through semantic evidence. Whereas extraction was based on orthographic similarity, classification was based on semantic similarity, considering the path length of the word pair in EuroWordNet and word co-occurrences in corpora.

In the last few years, static word embeddings have played a significant role in distinguishing false friends from cognates in a bilingual space, where cognates are expected to be closer in the space and false friends much more distant. In computational semantics, word embeddings, i.e., distributional vectors created with neural networks, are currently the dominant model to represent lexical meaning. Words are represented as real-valued numbers in vectors, with each number capturing a dimension of the meaning of each word, allowing semantically related words to be mapped to proximate points in the vector-space model. Indeed, word-vector models can be trained from two different approaches: count models and predictive models (Baroni et al., 2014). On the one hand, distributed semantic models can capture the importance of contexts by using linear algebra on word-to-word co-occurrence counts. On the other hand, predictive models, or neural-network models, use a non-linear function of word co-occurrences, where word embeddings capture more complex information than merely co-occurrence counts. According to Mandera et al. (2017), predictive models are much more psychologically grounded than count models because the underlying principle of implicitly learning how to predict a word from other words is consistent with biologically inspired models of associative learning.

Popular neural-network models for obtaining static word embeddings include Word2Vec (Mikolov et al., 2013), which contains a single hidden layer that, for each input word, returns the probability that the other words in the corpus belong to its context; GloVe (Pennington, et al., 2014), which builds word embeddings by considering the frequency of co-occurrences across the entire corpus; and FastText (Bojanowski et al., 2017), which enriches Word2Vec embeddings with sub-word information using bags of character n-grams. Static word embeddings have been lately explored to enhance deep-learning language models, leading to a breakthrough in natural language understanding (Periñán-Pascual, 2022). As of late, with the advent of Transformers such as BERT (Devlin et al., 2019) and other Large Language Models (LLMs), there is an open line of research into contextualised embeddings, which consider the linguistic meaning of nearby tokens in a given sentence.

Static word embeddings enhanced cognate recognition when cross-lingual word-embedding mapping was performed, initially used to generate bilingual dictionaries. These mapping methods, most of which are supervised and employ a small seed dictionary to learn the mapping (e.g., Mikolov et al., 2013; Smith et al., 2017), can align two sets of embeddings trained independently on monolingual corpora into a shared space. In this context, the challenge is to devise a fully unsupervised method so that parallel data from a bilingual dictionary or a sentence-aligned corpus is not required. To this end, Conneau et al. (2018) used adversarial learning to generate a rotation matrix, where the resulting space was then refined via Procrustes and its density was adjusted. Artetxe et al. (2018) combined an initial weak mapping that exploits the structure of the embedding spaces with a self-learning approach that, starting from the initial solution, iteratively improves the mapping. Wang, Henderson and Merlo (2019) proposed a weakly supervised adversarial-training method, where cross-lingual mapping was performed at the concept level through aligned Wikipedia articles rather than at the word level. Few studies have recently attempted to use word embeddings in the context of cognate recognition, as described below.

Castro et al. (2018) constructed vector spaces from the Spanish and Portuguese Wikipedia with Word2Vec and then used WordNet as the bilingual lexicon to align both vector spaces applying the technique described in Mikolov et al. (2013). Finally, they employed Support Vector Machines (SVMs) to classify word pairs as cognates or false friends. This supervised binary classifier was trained with several features based on the cosine distances between source and target vectors.

Merlo and Rodriguez (2019) explored two models, i.e., Artetxe et al. (2018) and Wang et al. (2019), to demonstrate that the structure of the cross-lingual word-embedding space yields the same similarity effects as the human bilingual lexicon.

Labat and Lefever (2019) applied SVM to identify cognates by combining orthographic similarity features from fifteen metrics with semantic information from word embeddings. Particularly, they employed FastText to get monolingual word embeddings for Dutch and English. After aligning them into a common vector space based on the model proposed by Smith et al. (2017), cosine similarity between the embeddings of the source and target words was computed and then introduced in the classifier. Lefever et al. (2020) extended this preliminary work in several aspects. For example, they incrementally re-trained the FastText embeddings with corpus-specific information, used a more advanced embedding mapping method, such as Artetxe et al. (2018), and combined orthographic and semantic features through a multi-layer perceptron.

Palmero Aprosio et al. (2020) used Word2Vec and FastText embeddings. First, candidate cognates were extracted from a corpus with the help of a monolingual dictionary in another language and several orthographic similarity metrics. Second, candidates were recognised as true cognates or false friends by an SVM classifier, trained with semantic features obtained from aligned multilingual spaces, synonym lists, and bilingual dictionaries.

Uban et al. (2019) employed the publicly available multilingual aligned matrices constructed by Smith et al. (2017).[11] In the mapping process of 78 languages to a single common space, Smith et al. (2017) aligned all monolingual FastText word embeddings directly to the English vectors, so they did not rely on training dictionaries between non-English language pairs. Uban et al. (2019) quantified the semantic divergence of cognate word pairs through the cosine distance between source and target words in this shared embedding space so that the degree of falseness was measured as follows in Equation (1):

(1) Degree of falseness = distance(c_1, w_2) – distance(c_1, c_2)

In this equation, (c_1, c_2) is a cognate pair, i.e., c_1 is a word in the source language and c_2 is in the target language. If a word w_2 in the target language is semantically closer to c_1 than c_2 in the shared semantic space, the cognates can be considered false friends.

Finally, Uban and Dinu (2020) used the same procedure described above but with FastText multilingual pre-aligned embeddings constructed by Conneau et al. (2018).[2]

3. Proposed method

Our study was influenced by Taylor (2012), who explored the notions of E-language (external language) and I-language (internal language). On the one hand, a language can be seen as an "external" object, i.e., a set of linguistic realisations, an approach endorsed by Bloomfield (1926). On the other hand, a language can be considered an "internal" object, i.e., lexical and grammatical knowledge that resides in speakers' minds, a cognitive turn initiated by Chomsky's (1957) generative model. Taylor (2012) supported the idea that both models are closely aligned. This distinction of linguistic objects can result in the construction of two types of language resources: text corpora, which exemplify the external approach, and lexical associative networks, which represent the internal approach.

In psycholinguistics, De Deyne et al. (2016) suggested that, when people judge semantic associations, they tend to rely more on lexical networks than on the distributional properties of an external language model. For example, *yellow* and *banana* rarely co-occur in text corpora because most bananas are yellow, so mentioning yellow together with banana is uninformative. However, *yellow* and *banana* are strongly associated in WordNet as shown in their high semantic similarity. In this context, this study aims to compare the capacity of three types of knowledge (i.e., corpus-based knowledge, semantic knowledge, and commonsense knowledge) to detect true cognates and false friends in English and Spanish. To this end, we leveraged pre-trained word embeddings from different types of knowledge sources so that our research method can be easily applied to other languages.

On the one hand, we employed FastText embeddings (Bojanowski et al., 2017) to analyse semantic similarity in a corpus-based model, which is grounded on the knowledge derived from the distributional patterns of words. In our experiment, we employed publicly available FastText embedding matrices of English and Spanish, each one containing 2M tokens and 300 dimensions trained on Wikipedia and the Common Crawl corpus, which contains petabytes of data collected in fifteen years of web crawling.[3] The models were trained using CBOW with character n-grams of length 5, a window of size 5 and 10 negatives. FastText learns word representations by considering sub-word information. In particular, this method incorporates character n-grams into the Skip-gram architecture of Word2Vec, where the representation of words finally results from the sum of the n-gram vectors. As this model was trained with a collection of textual expressions, it reflects the E-language of speakers.

On the other hand, the network-based model aims to construct word embeddings from the semantic structure of knowledge bases without any corpus evidence, where relational information is exploited by graph-based methods, considering word senses as nodes and semantic relations between senses as edges. To this end, we constructed WordNet embeddings with the Wnet2vec method (Saedi et al., 2018), using the English WordNet 3.0 (Fellbaum, 2010)[4] and the MCR Spanish WordNet 3.0 datasets (Agire et al., 2012).[5] As human experts constructed WordNet relying on existing lexicographic sources and introspection, it reflects the I-language of speakers. Saedi et al. (2018) proposed an effective method to re-encode semantic networks into word-embedding matrices based on the intuition that the larger the number of paths and the shorter the paths connecting two nodes in the ontological graph, the stronger their semantic affinity. In this conversion, different semantic relations are considered among all parts of speech, where identical weight is given to each relation. Moreover, Positive Pointwise Mutual Information is used to reduce the bias introduced by highly polysemous words. After each vector is L2-normalised, Principal Component Analysis reduces the size of the vectors. Saedi et al. (2018, p. 129) reported that "the performance of wnet2vec was around 15% superior to the performance of word2vec" when "evaluated under the mainstream task of determining the semantic similarity of words arranged in pairs", thus demonstrating that some semantically related words do not co-occur in relevant contexts.

Finally, hybrid models combine the knowledge in corpus-based distributional representations and the relational information extracted from knowledge bases. In this regard, LessLex[6] (Colla et al., 2020) was constructed from two primary resources: BabelNet and ConceptNet Numberbatch. On the one hand, BabelNet (Navigli & Ponzetto, 2012) is a semantic network resulting from the integration of lexicographic and encyclopaedic knowledge from WordNet and Wikipedia, where word senses are represented as synsets. On the other hand, ConceptNet Numberbatch (Speer et al., 2017) was built through an ensemble method combining the embeddings produced by GloVe and Word2vec with the structured knowledge from the semantic networks ConceptNet (Havasi et al., 2007) and the Paraphrase Database (Ganitkevitch et al., 2013). ConceptNet Numberbatch word embeddings were taken as the starting point, as vectors are mapped onto a single shared multilingual semantic space covering over seventy-eight different languages. Averaging vectors from the above resources, LessLex finally contains embeddings for senses rather than for terms, allowing us to compute the distance between a term and each of its corresponding senses. Therefore, LessLex integrates corpus-based knowledge from Wikipedia, semantic knowledge from WordNet, and commonsense knowledge from ConceptNet. It should be noted that, in our experiment, we did not only leverage LessLex sense embeddings, each corresponding to a BabelNet ID, but also converted them into word embeddings for English and Spanish. In the case of polysemous words, i.e., a word in a given language linked to more than one BabelNet ID, word embeddings were computed from the average of the vectors involved in the different senses.

Moreover, we assessed the impact of cross-lingual embedding alignment in semantic similarity discovery, whereby English and Spanish matrices were projected onto a shared space in an unsupervised fashion following Artetxe et al.'s method (2018). In this way, two monolingual matrices projected onto the same space can be merged into a single bilingual word-embedding matrix.

Table 1 summarises the experiment conducted in this study, where we evaluated six different models through nine word-embedding matrices to automatically recognise true cognates and false friends.

Table 1

Word-embedding models employed in the experiment.

Model type	**Unaligned model**	**Aligned model**
Corpus-based model	Two monolingual FastText matrices (Model I)	One bilingual FastText matrix (Model IV)
Network-based model	Two monolingual Wnet2vec matrices (Model II)	One bilingual Wnet2vec matrix (Model V)
Hybrid model	Two monolingual LessLex matrices (Model III)	One bilingual LessLex matrix (Model VI)

The basic premise of the above models is that lexical meaning results from the sum of word senses, as the experiment is conducted with out-of-context words. In this regard, an issue often discussed in relation to word-embedding models is what Camacho–Collados and Pilehvar (2018) called "meaning conflation deficiency". Specficically, such models do not discriminate among different senses of a word, as each word type has a single word vector, so polysemy and homonymy are ignored. As a result, multi-sense embedding models are constructed to deal with the meaning conflation deficiency of word embeddings (Iacobacci et al., 2015; Ruas et al., 2019). However, Kober et al. (2017) demonstrated that a single vector that conflates the different senses of a polysemous word is sufficient for recovering sense-specific information and thus discriminating the meaning of a word in context in tasks such as word-sense disambiguation. For this reason, we also aimed to determine whether an independent representation for each word sense in a vector-space model could contribute to improving the recognition of true cognates and false friends. To this end, we employed LessLex sense embeddings (Model VII) in contrast to monolingual and bilingual matrices in Model III and Model VI, which are grounded on LessLex word embeddings.

4. Experiment

The purpose of the experiment was to assess the capability of the seven models described in Section 3 to automatically detect true cognates and false friends in a test dataset. To achieve this, we created the test dataset as a gold standard, taken as the ground truth, from two lists of cognate word pairs in English and Spanish (i.e., a list of 3,861 true cognates and another of 430 false friends) compiled for Foreign Language Teaching by Rubén Morán.[7] Our test dataset consisted of 1,164 English-Spanish word pairs, including 582 orthographically similar word pairs (i.e., 291 true cognates and 291 false friends) and 582 word pairs randomly created from the vocabulary involved in the true cognates and false friends. The selection of such cognates from the original lists was based on the criteria of balance and coverage, i.e., (a) a balanced proportion not only between cognates and non-cognates but also between true cognates and false friends, and (b) the coverage of all selected words in the FastText, Wnet2vec and LessLex word-embedding matrices. Moreover, the original matrices were finally preprocessed to facilitate the comparison, resulting in 300-dimensional matrices that only contained the word embeddings corresponding to 575 English words and 563 Spanish words in the test dataset. Accordingly, bilingual aligned matrices consisted of 1,138 vectors.

The experiment was conducted in two stages. First, we performed the orthographic similarity (OS) analysis of word pairs, which was based on the Normalised Levenshtein Distance (NLD) between the source and target words in each pair in the test dataset, as in Equation (2).

(2) $\text{OS}=1\text{-NLD, where NLD}=\frac{\text{distance}}{\text{length}}$,

distance = min(*number of insertions, deletions and substitutions*), and

length = max(*length of source word, length of target word*)

In NLD, the Levenshtein Distance, which represents the minimum number of insertions, deletions, and substitutions of one character for another when transforming one word into the other, is divided by the maximum length of both words. As OS is a symmetric measure, the source word can be in English or Spanish, irrespectively. The goal of this stage was to determine the optimal threshold above which the metric yielded the best performance based on binary classification. In other words, we used OS scores as threshold values, ranging from 0 to 1 in increments of 0.05. Then, we determined whether each word pair in the test dataset could be classified as cognate or non-cognate according to a given threshold.

With respect to evaluation metrics for binary classification, it should be recalled that most of them are built over a 2x2 contingency matrix—as shown in Table 2, where TP, FP, FN and TN denote the number of true positives, false positives, false negatives, and true negatives, respectively.

Table 2

Contingency matrix for binary classification.

			Expected (Ground Truth)	
			Are the words in the pair really cognates?	
			yes	no
Predicted	Were the words in the pair classified as cognates?	yes	TP	FP
		no	FN	TN

Using the information from this contingency matrix, we employed the popular measure of Accuracy, which represents the proportion of correctly classified samples (i.e., word pairs) from the total number of samples (i.e., both positive and negative classes), as in Equation (3).

$$(3)\ \text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Given the relative balance of our test dataset, Accuracy could be an appropriate metric. However, it does not guarantee an adequate assessment of the overall effectiveness of the model, as the accuracy paradox reveals that "high accuracy is not necessarily an indicator of high classifier performance" (Valverde-Albacete & Peláez-Moreno, 2014, p. 2). Indeed, it is misleading to interpret results as accurate when the impact of errors (i.e., false positive and false negative) is minimised, even though they are critical factors to be considered when evaluating classifiers.

Accordingly, we also employed typical evaluation metrics that come from information retrieval, such as Precision (Equation 4), Recall (Equation 5), and F1 (Equation 6).

$$(4)\ \text{Precision}=\frac{TP}{TP+FP}$$

$$(5)\ \text{Recall}=\text{Sensitivity}=\frac{TP}{TP+FN}$$

$$(6)\ \text{F1}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

Precision represents the proportion of correctly classified positive samples to the total number of samples classified as positive, either correctly or incorrectly. Therefore, the higher the Precision, the fewer incorrect positive classifications. Recall (sensitivity) represents the ratio of correctly classified positive samples to the total number of positive samples. Therefore, the higher the Recall, the more positive samples recognised. F1 represents the harmonic mean of Precision and Recall. Therefore, F1 estimates the relative impact of the errors caused by FPs (Precision) and the errors caused by FNs (Recall) on the model under evaluation. After determining the optimal threshold based on F1, this stage returned a list of word pairs whose OS scores were above a particular cutoff.

Second, we performed the semantic similarity (SS) analysis of word pairs based on the seven models, where a cosine similarity score was computed for each candidate recognised in the previous stage. We used SS scores as threshold values, ranging from 0 to 1 in increments of 0.05. The goal of this stage was to determine the optimal threshold above which SS yielded the best performance based on Accuracy, Precision, Recall, and F1.

However, F1 is not an efficient metric to test model effectiveness if we also intend to consider TNs. In this regard, Specificity (Equation 7) is a relevant measure which expresses the ratio of correctly classified negative samples to the total number of negative samples. Moreover, we employed three popular measures that combine Sensitivity (Recall) and Specificity, giving a comprehensive evaluation of the overall performance of the model: Youden's Index (Equation 8), Likelihood Ratio (Equation 9 and 10), and Diagnostic Odds Ratio (Equation 11), which originated in medical diagnosis to analyse tests.

(7) $\text{Specificity}=\frac{\text{TN}}{\text{FP+TN}}$

(8) $\text{YI}=\text{Sensitivity}+\text{Specificity}-1$

(9) $\text{LR+}=\frac{\text{Sensitivity}}{\text{1-Specificity}}$

(10) $\text{LR-}=\frac{\text{1-Sensitivity}}{\text{Specificity}}$

(11) $\text{DOR}=\frac{\text{LR+}}{\text{LR-}}$

Youden's Index (YI) gives equal weight to FPs and FNs. Therefore, when two models have the same YI, they have the same proportion of total misclassified samples. A high value of YI indicates a better ability to avoid failure, so 1 indicates that there are no FPs or FNs, and thus the classifier is perfect. Likelihood Ratio (LR) represents how likely samples are assigned a given category. The higher the value, which ranges from 0 to infinity, the more likely they have the category. In particular, Positive Likelihood Ratio (LR+) and Negative Likelihood Ratio (LR-) represent the extent to which positive and negative predictions, respectively, indicate the outcome will be positive. Therefore, high LR+ and LR- mean better performance in positive and negative classes, respectively. For example, if LR+ is 0.7, the model is moderately good to very good at classifying the positive class. That is, 0.7 indicates that positive prediction is 0.7 times more likely to be positive cases than negative cases. Conversely, if LR- is 0.3, it is less helpful when ruling out negative cases. That is, 0.3 indicates negative predictions are 0.3 times as likely to be positive cases than negative cases. LR+ and LR- are combined into one measure, i.e., Diagnostic Odds Ratio (DOR), which represents the ratio between LR+ and LR-. The higher the DOR value, which ranges from 0 to infinity, the more indicative of good performance. Therefore, a value greater than 1 means that the classifier is valid, 1 means that the model does not provide useful information, and a value less than 1 indicates that the model predicts in the wrong direction. In conclusion, these three measures were critical when comparing the models and determining which model is preferable.

However, no single measure of the above is grounded on the four values in the confusion matrix. For this reason, we also applied Matthews Correlation Coefficient (MCC), which computes a high score only if the classifier correctly predicts most of the positive samples and most of the negative samples, and if most of its positive predictions and most of its negative predictions are also correct. MCC measures the correlation of the expected classes (i.e., ground truth) with the predicted classes, as in Equation (12).

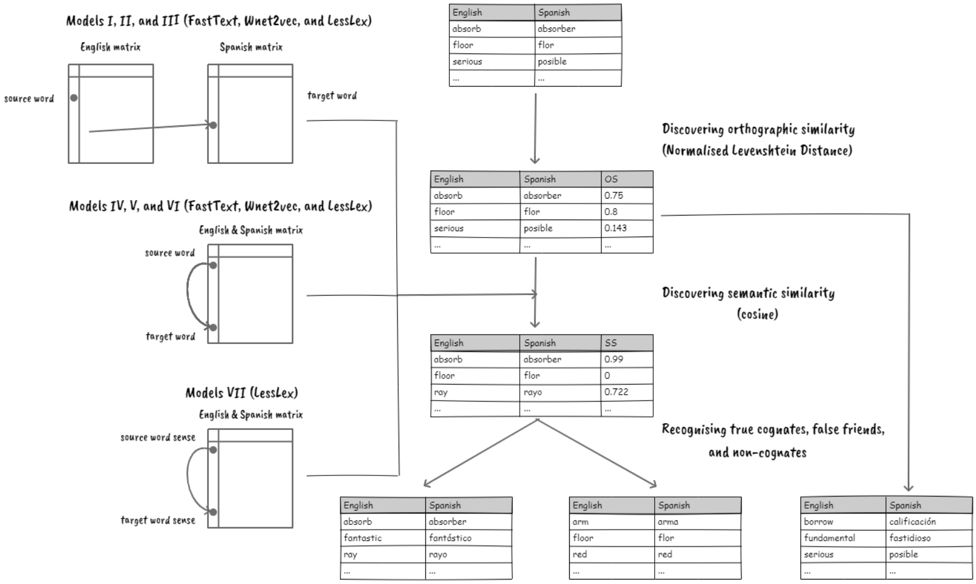
$$(12)\ MCC=\frac{TP\times TN\text{-}FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

MCC returns a value from -1 to +1, where +1 describes a perfect prediction, 0 represents a random prediction, and -1 describes a perfectly wrong prediction (i.e., predicted and expected values completely disagree). Recent scientific studies (Chicco & Jurman, 2020; Chicco et al., 2021a; Chicco et al., 2021b) have shown that MCC is more reliable than Accuracy, F1, Youden's Index, and DOR in binary classification evaluation. Except for LR, DOR and MCC, the other evaluation measures compute a score ranging from 0 to 1, where 0 is a poor result, and 1 represents the perfect outcome.

To conclude, Figure 1 illustrates the process of cognate detection and classification.

Figure 1

Cognate detection and classification.



The pipeline in Figure 1 works as follows: first, the task of cognate detection is performed with an unsupervised orthographic approach based on OS between a given pair of words through the NLD. Then, after finding the best threshold, the task of cognate classification is carried out for orthographically similar words to distinguish true cognates from false friends (i.e., partial, false, and deceptive cognates). Cognate classification is performed with the semantic similarity measure using different unsupervised embedding-based models, selecting the threshold that gives the highest performance for each approach.

5. Results

In the first stage of the experiment, we performed cognate detection to evaluate the capability of the system to differentiate between cognates and non-cognates based on OS (Model 0). In the second stage of the experiment, we selected the word pairs recognised as orthographically similar in the first stage, particularly those obtained from the threshold with the best performance (i.e., cutoff 0.4), to perform cognate classification. In other words, we then evaluated the capability of the system to differentiate between true cognates and false friends based on SS from various models. Table 3 presents the best evaluation results for each model with respect to F1, where we point out the highest F1 score in bold. We also tested other aspects of model effectiveness to complement the above results. Table 4 shows the scores computed by YI, LR+, LR-, DOR, and MCC for the same cutoffs presented in Table 3.

Table 3

Best evaluation results.

Task	Model	Cutoff	Accuracy	Precision	Recall	F1
Cognate detection	**0**	0.4	0.984	0.980	0.988	0.984
	I	0	0.486	0.486	1	0.654
	II	0	0.486	0.486	1	0.654
	III	0.95	0.856	0.825	0.789	0.807
Cognate classification	**IV**	0.55	0.813	0.776	0.863	**0.817**
	V	0	0.486	0.486	1	0.654
	VI	0.8	0.777	0.719	0.888	0.794
	VII	1	0.722	0.641	0.975	0.773

Table 4

Best evaluation results [continued].

Task	Model	YI	LR+	LR-	DOR	MCC
Cognate detection	**0**	0.967	47.961	0.012	3996.750	0.967
	I	0	1	-	-	-
	II	0	1	-	-	-
	III	0.687	7.692	0.235	32.732	0.693
Cognate classification	**IV**	0.628	3.672	0.179	20.514	0.630
	V	0	1	-	-	-
	VI	0.560	2.708	0.167	16.216	0.571
	VII	0.459	1.888	0.051	37.02	0.522

6. Discussion

The first stage of the experiment, related to cognate detection, demonstrated that the most reliable list of orthographically similar word pairs was automatically constructed with a threshold of 0.4. Indeed, the orthographic approach outperformed the semantic approach in the nine metrics, where not only Accuracy and F1 (i.e., 0.984), but also YI and MCC (i.e., 0.967) revealed that the classifier is close to being perfect. The second stage of the experiment, which explored cognate classification, demonstrated the impact of matrix projection together with the type of knowledge to recognise true cognates and false friends. On the one hand, we discovered that working with unaligned matrices does not tend to provide good results, as shown by the Accuracy scores obtained with Model I and Model II. Table 3 shows that the highest F1 in Model I and Model II (0.654 with a threshold of 0) is the lowest among all models. Considering this cutoff in both models, which produces the lowest Accuracy (0.486), Table 4 shows that YI is 0, and LR-, DOR and MCC cannot even be calculated, so these classifiers are not valid. This evidence proves that FastText and Wnet2vec unaligned matrices are not useful for the cognate classification task. We attribute this inefficiency to the lack of alignment of the matrices involved, as semantic representations are compared from different vector spaces. However, Model III, which is supposed to have no cross-lingual word-embedding mapping, is one of the best models. The issue is that LessLex word embeddings were generated from sense embeddings, so Model III is inherently aligned as semantic representations of translation equivalents in two different languages (i.e., words of the same WordNet synsets) originated from the same conceptual representations.

On the other hand, we discovered that aligned embeddings could improve evaluation results only when vectors were derived from corpora. Indeed, the F1 score of Model IV (FastText) is the highest among all models (Table 3). Moreover, model effectiveness metrics such as YI, LR+, LR-, DOR, and MCC revealed that the projection to a shared space proved beneficial (Table 4). However, the evaluation results of Model V (i.e., aligned Wnet2vec matrix) are similar to those of Model II (i.e., unaligned Wnet2vec matrices), where not only Accuracy and F1 (Table 3), but also YI (Table 4) revealed that Model V is not suitable for our task. Therefore, we conclude that models based on semantic networks (e.g., WordNet) do not contain sufficient information to recognise true cognates and false friends. In the case of Model VI, cross-lingual projection did not produce better results with LessLex word embeddings, i.e., 0.794 in F1 (cutoff 0.8). Although such results are reasonably good, all metrics on model effectiveness revealed that Model VI is not superior to III (Table 4). Again, this supports the idea that we tried to align vectors within an already shared semantic space.

Moreover, we closely examined the models with the best performance in the second stage of the experiment: Model III, with two unaligned monolingual matrices based on corpus-based, semantic, and commonsense knowledge, and Model IV, with a single matrix that contains aligned bilingual embeddings constructed from corpora. On the one hand, Model III outperforms Model IV and all the other models in this stage in terms of Accuracy and Precision (0.856 and 0.825, respectively), where the optimal threshold of 0.95 contributes to providing the lowest Recall (0.789). On the other hand, Model IV is superior to Model III in Recall and F1 (0.863 and 0.817, respectively), where the optimal threshold of 0.55 contributes to improving Recall at the expense of Precision (0.776). Therefore, only after examining the model effectiveness metrics in Table 4, we were able to establish the best model, revealing that Model III yielded the highest YI, DOR, and MCC scores. Moreover, Model III produced the highest LR+ and LR-, which implies that this model is superior for confirming positive samples, being in line with the goal of our task.

Finally, LessLex word embeddings (Model III and Model VI) outperformed LessLex sense embeddings (Model VII) in terms of Accuracy, Precision, F1, i.e., 0.722, 0.641, and 0.773, respectively. Moreover, Model VII was one of the least effective models, as shown by YI, LR+, LR-, and MCC (i.e., 0.459, 1.888, 0.051, and 0.522, respectively). This shows that independent representation for each word sense in a vector-space model could not improve the recognition of true cognates and false friends, a task performed out of context. Therefore, we conclude that meaning conflation deficiency did not affect performance.

CONCLUSIONS

True cognates and false friends have sparked research interest in fields such as applied linguistics and NLP. Due to their ever-growing importance and the time-consuming and labour-intensive task of manually curating lists of cognates and false friends, it becomes essential to develop automatic means to compile and evaluate such lists. Thus, our contribution is to fill this gap by assessing the impact of three aspects of unsupervised embedding-based models (i.e., language resource type, cross-lingual embedding alignment, and meaning conflation) in the semantic-similarity stage of the cognate classification task of distinguishing true cognates from false friends with English and Spanish words. The results of our experiments indicate that (i) the best model is the LessLex word-embedding model, based on corpus-based, semantic, and commonsense knowledge, which does not benefit from cross-lingual embedding alignment because it already contains built-in multilingual knowledge; (ii) cross-lingual matrix embedding alignment significantly improved the performance of our corpus-based word-embedding model (i.e., FastText word-embedding model) in the cognate classification task but did not offer any improvement in the network-based model (i.e., WordNet word-embedding model); and (iii) meaning conflation deficiency did not negatively affect the performance of any of the word-embedding models. These results support the excellent predictive capabilities of fully unsupervised approaches based on word embeddings. Our experiment could provide a basis for future research in cross-lingual alignment tasks related to NLP, such as semantic similarity or machine translation, or in foreign language learning, by helping to construct lists of cognates automatically. Other lines of research could focus on the evaluation of lists of cognates and false friends for pairs of languages other than English and Spanish to investigate whether the same results may apply by using a multilingual database of cognates, e.g., the CogNet database (Batsuren et al., 2022). We also plan to make use again of novel NLP resources, such as monolingual transformer-based embeddings like BERT (Devlin et al., 2019) and multilingual transformer-based embeddings like XLM-RoBERTa (Barbieri et al., 2021), which have shown promising results in semantic similarity tasks with respect to more traditional methods (Chandrasekaran & Mago, 2022). Finally, we intend to employ advanced techniques in prompt engineering with open-source LLMs for cognate classification, saving computational costs and time, and thus addressing ecological and cost-efficiency concerns.

REFERENCES

Agirre, A. G., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0: Upgrading a very large lexical knowledge base. In *GWC 2012: 6th International Global Wordnet Conference, Proceedings* (pp. 118-125).

Artetxe, M., Labaka, G., & 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* (pp. 789-798). Association for Computational Linguistics.

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2021). XLM-T: A Multilingual Language Model Toolkit for Twitter. http://arxiv.org/abs/2104.12250

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238-247). Association for Computational Linguistics.

Batsuren, K., Bella, G., & Giunchiglia, F. (2022). A large and evolving cognate database. *Language Resources and Evaluation, 56*(1), 165-189.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2*,* 153-64.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics,* 5, 135-146.

Brew, C., & McKelvie, D. (1996). Word-pair extraction for lexicography. In *Proceedings of the second international conference on new methods in language processing* (pp. 45-55). Ankara, Turkey.

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research,* 63, 743-788.

Castro, S., Bonanata, J., & Rosá, A. (2018). A high coverage method for automatic false friends detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 29-36).

Chandrasekaran, D., & Mago, V. (2022). Evolution of semantic similarity: A survey. *ACM Computing Surveys, 54*(2), 1-37.

Chicco, D., & Jurma, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 1-13.

Chicco, D., Starovoitov, V., & Jurman, G. (2021a). The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access, 9*, 47112-47124.

Chicco, D., Tötsch, N., & Jurman, G. (2021b). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining, 14*(13), 1-22.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Colla, D., Mensa, E., & Radicioni, D. P. (2020). LESSLEX: Linking multilingual Embeddings to SenSe representations of LEXical items. *Computational Linguistics, 46*(2), 289-333.

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *Proceedings of the Sixth International Conference on Learning Representations* (pp. 1-14).

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th international conference on computational linguistics* (pp. 1861-1870).

De Groot, A. M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals: An Introduction*. Psychology Press.

De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1-56.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 4171-4186).

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*(3), 284-30.

Dominguez, P. J. C., & Nerlich, B. (2002). False friends: Their origin and semantics in some selected languages. *Journal of Pragmatics, 34*(12), 1833-1849.

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231-243). Springer Netherlands.

Frunza, O., & 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* (pp. 91-100).

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT* (pp. 758-764). Atlanta, GA.

Granger, S. (1993). Cognates: An aid or a barrier to successful L2 vocabulary development? *ITL Review of Applied Linguistics*, 99-100, 43-56.

Gurrutxaga, A., Saralegi, X., Ugartetxea, S., & Alegria, I. (2006). Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *Atti del XII Congresso Internazionale di Lessicografia* (pp. 159-165).

Hansen-Schirra, S., Nitzke, J., & Oster, K. (2017). Predicting cognate translation. In S. Hansen-Schirra, O. Czulo, & S. Hofmann (Eds.), *Empirical modelling of translation and interpreting* (pp. 3-22). Language Science Press.

Havasi, C., Speer, R., & Alonso, J. (2007). ConceptNet: A lexical resource for common sense knowledge. In N. Nicolov, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing V: Selected Papers from RANLP* (pp. 269-280). Borovets, Bulgaria.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2015). Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 95-105).

Inkpen, D., & Frunza, O. (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' 05)* (pp. 251-257). Borovets, Bulgaria.

Kober, T., Weeds, J., Wilkie, J., Reffin, J., & Weir, D. (2017). One Representation per Word - Does it make Sense for Composition? In J. Camacho-Collados & M. T. Pilehvar (Eds.), *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications* (pp. 79-90).

Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 288-295). Seattle, WA.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In *Proceedings of the 2nd meeting of the North American chapter of the association for computational linguistics (NAACL 2001)* (pp. 103-110). Pittsburgh, PA.

Kondrak, G., Marcu, D., & Knight, K. (2003). Cognates can improve statistical translation models. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*.

Labat, S., & Lefever, E. (2019). A classification-based approach to cognate detection combining orthographic and semantic similarity information. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 602-610).

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

Lefever, E., Labat, S., & Sing, P. (2020). Identifying cognates in English-Dutch and French-Dutch by means of orthographic information and cross-lingual word embeddings. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, European Language Resources Association (pp. 4096-4101).

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language,* 92, 57-78.

Melamed, D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics, 25*(1), 107-130.

Merlo, P., & Rodriguez, M. A. (2019). Cross-lingual word embeddings and the structure of the human bilingual lexicon. In *Proceedings of the 23rd Conference on Computational Natural Language Learning* (pp. 110-120).

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. https://arxiv.org/abs/1309.4168

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience,* 40, 543-566.

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

Mulloni, A., Pekar, V., Mitkov, R., & Blagoev, D. (2007). Semantic evidence for automatic identification of cognates. In *Proceedings of the 1st international workshop on acquisition and management of multilingual lexicons* (pp. 49-54). Borovets, Bulgaria.

Nakov, S., Nakov, P., & Paskaleva, E. (2007). Cognate or false friend? Ask the web! In *Proceedings of the 1st International Workshop on Acquisition and Management of Multilingual Lexicons, held in conjunction with Recent Advances in NLP 2007* (pp. 55-62).

Navigli, R., & Ponzetto, S. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence,* 193*,* 217-250.

Nazar, R. (2011). Parallel corpus alignment at the document, sentence and vocabulary levels. *Procesamiento del Lenguaje Natural,* 47, 129-136.

Ng, E.-L., Chin, B., Yeo, A., & Ranaivo-Malancon, B. (2010). Identification of closely-related indigenous languages: An orthographic approach. *International Journal on Asian Language Processing, 20*(2), 43-61.

Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-linguistic similarity matter? *Language Learning, 70*(3), 685-731.

Palmero Aprosio, A., Menini, S., & Tonelli, S. (2020). Adaptive complex word identification through false friend detection. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalisation* (pp. 192-200).

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532-1543).

Periñán-Pascual, C. (2022). Measuring associational thinking through word embeddings. *Artificial Intelligence Review*, 55, 2065-2102.

Procter, P. (1995). *Cambridge International Dictionary of English*. Cambridge University Press.

Ruas, T., Grosky, W., & Aizawa, A. (2019). Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications,* 136, 288-303.

Sabino, M. A. (2002). *Dicionário de Falsos Cognatos e Cognatos Enganosos Italiano-Português: Subsídios Teóricos e Práticos* [Tesis de doctorado]. Araraquara, FCL.

Sabino, M. A. (2016). False cognates and deceptive cognates: Issues to build special dictionaries. In *Proceedings of the 17th EURALEX International Congress* (pp. 746-755).

Saedi, C., Branco, A., Rodrigues, J. A., & Silva, J. (2018). WordNet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP* (pp. 122-131). Melbourne, Australia.

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*(1), 157-166.

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations.*

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 4444-4451).

Taylor, J. R. (2012). *The Mental Corpus: How Language is Represented in the Mind.* Oxford University Press.

Uban, A.-S., Ciobanu, A., & Din, L. P. (2019). A computational approach to measuring the semantic divergence of cognates. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 1-12).

Uban, A.-S., & Dinu, L. P. (2020). Automatically building a multilingual lexicon of false friends with no supervision. In *Proceedings of the 12th Conference on Language Resources and Evaluation, European Language Resources Association* (pp. 3001-3007).

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PloS One, 9*(1), 1-10.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities, 32*(2/3), 73-89.

Wang, H., Henderson, J., & Merlo, P. (2019). Weakly-supervised concept-based adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association of Computational Linguistics* (pp. 4419-4430). Hong Kong, China.

ACKNOWLEDGEMENTS

This publication is part of the R&D&I project PID2020-112827GB-I00, funded by MICIU/AEI/10.13039/501100011033.

NOTES

[1] https://github.com/babylonhealth/fastText_multilingual

[2] https://github.com/facebookresearch/MUSE

[3] https://fasttext.cc/docs/en/crawl-vectors.html

[4] https://wordnet.princeton.edu/download/current-version

[5] http://adimen.si.ehu.es/web/MCR/

[6] https://ls.di.unito.it/resources/lesslex/

[7] http://www.cognates.org/pdf/mfcogn.pdf,
http://www.cognates.org/pdf/false-cognates.pdf