Friend or Foe? A Computational Investigation of Semantic False Friends across Romance Languages

Ana Sabina Uban[♠,♡] **Liviu P. Dinu**[♠,♡] **Bogdan Iordache**[♠,♡]
Simona Georgescu[♣,♡] **Claudia Vlad**[♣,♡]
University of Bucharest, [♠] Faculty of Mathematics and Computer Science,
[♣]Faculty of Foreign Languages and Literatures, [♡]HLT Research Center
{ldinu, auban}@fmi.unibuc.ro, ioan.iordache@s.unibuc.ro,
{simona.georgescu, claudia.vlad}@lls.unibuc.ro

Abstract

In this paper we present a comprehensive analysis of lexical semantic divergence between cognate words and borrowings in the Romance languages. We experiment with different algorithms for false friend detection including deceptive cognate and deceptive borrowings and correction and evaluate them systematically on cognate and borrowing pairs in the five Romance languages. We use the most complete and reliable dataset of cognate words and borrowings based on etymological dictionaries for the five main Romance languages (Italian, Spanish, Portuguese, French and Romanian) to extract deceptive cognates and borrowings automatically based on usage, and freely publish the lexicon of obtained true and deceptive cognate and borrowings in every Romance language pair.

1 Introduction and Related Work

Words are the interface of a language, the first layer with which one comes into contact when approaching that language. If it is a language related to the speaker's mother tongue, or to another language he or she knows, the amount of words similar to the known language will be the first aid to learning. These similar words in language A and language B are, in most cases, cognates (words that derive from the same etymon in language C, e.g. Ro. *casa*, Es. *casa* 'house' are cognates because they both come from La. *casa* 'hut'), or, in fewer cases, borrowings (words adopted in language A - target language - from language B - source language - or viceversa, e.g. Es. *valija* 'luggage' is a borrowing from It. *valigia* 'id.'). At first glance, these similarities seem a very promising environment for picking up a new foreign language. In reality, things are a bit more complicated.

Cognates inherited from Latin often have sufficiently different forms that their common origin is no longer transparent to speakers (e.g., Fr. *chien* - Ro. *câine* 'dog'). There is, however, a category of cognates whose form is very similar from one language to another, but which have different meanings (e.g. Es. *pariente* 'relative'- Ro. *părinte* 'parent'). This is what we call *deceptive cognates*, or, as Dominguez and Nerlich (2002) propose, *semantic false friends*: words whose meaning has diverged from that of the etymon and, consequently, from the meaning of cognates that have remained closer to the original concept. In addition to deceptive cognates, we have to distinguish the category of borrowings that no longer share the same meaning in the source and target language (e.g. Fr. *caracole* 'succession of right and left voltes executed by a horse' (target language) and Es. *caracol* 'snail' (source language), or Es. *novela* 'novel' (target language) and It. *novella* 'short story' (source language)). We will call these pairs *deceptive borrowings*, but, together with deceptive cognates, they may be referred to as semantic false friends. Dominguez and Nerlich (2002) distinguish them from *chance false friends*, which have similar forms and different meanings, but, unlike semantic false friends, do not share a common origin (e.g. Es. *nuca* 'nape' < Arabic *nuhā* 'marrow' vs. Ro. *nuca* 'nut' < La. *nux* 'id.').

In any of the cases cited above, the formal resemblance becomes misleading and inevitably leads the nonnative speaker of language A to tend to use them in the same contexts in which they appear in his or her mother tongue B (i.e., translating the term *w* in A by the term *w'* in B), distorting the information and making communication more difficult. That is why we propose a method for automatic detection. In the long run, we are interested in the circumstances that lead to semantic divergence of words stemming from the same etymon and, further on, the possibility of identifying recurrent trajectories in semantic change.

We agree that Romance words, in their evolution from Latin, have changed form. As for their

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15310–15324

meaning, a simple glance at Meyer-Lübke (1911) *Romanisches Etymologisches Wörterbuch* shows us that the faithful preservation of the Latin meaning occurs only in the case of a few concrete notions in domains where little perceptual change intervenes: body parts, family, common animals, etc. Words evolve with society, and semantic change occurs naturally, either from a need to designate new concepts (e.g. La. *dialis* 'day-light clock' > En. *dial* 'to telephone'), or expressive needs (e.g. En. *awfully* came to be used in positive contexts for the expressive impact it produced).

In most cases, the semantic change a > b is the result of an initial polysemy of a word. Take the example of La. *vindicare* 'to take revenge' vs. Ro. *vindeca* 'to cure'. This seems an utterly bizarre and incomprehensible change, unless we consider its full polysemy in Latin (the source language): 1. to claim ownership; 2. to claim; 3. *vindicare aliquem in libertatem* 'to call one to liberty, to set free'; 4. to set free, escape, save; 5. to punish, to avenge. The only meaning that has been inherited in Romanian is m4, with a medical specialization. At the same time, La. *vindicare* was preserved in Es. *vengar* with the same meaning of 'to take revenge', which creates a deceptive cognate pair Ro.-Es.

There are varying degrees of semantic divergence in a pair of false friends: in some cases there is a difference of connotation in the pragmatic register (e.g. It. *amico* 'friend' - Ro. *amic* 'acquaintance', and *prieten*, a word of Slavic origin is used for the concept of 'friend'), while in other situations cognates have completely different meanings (e.g. Ro. *larg* 'wide' vs Es. *largo* 'long' - in Es. the equivalent of Ro. *larg* is *ancho*, and the Romanian equivalent for Es. *largo* is *lung*).

The question that naturally arises is from what degree of semantic divergence can we talk about false friends? Within what distance is it permissible to say that it is just a different nuance, and the cognate can be used in context without distorting the message, and what distance is the sign of a divergence that is already impossible to overcome in the language?

Lexical semantic change has been studied with automatic methods based on distributed representations since the popularization of the first static word embeddings. The first study used static word embeddings trained on diachronic corpora (Hamilton et al., 2016), and more studies have elaborated on this research on different languages and for different tasks, (including for example, bias detection in embedding vector spaces, or graded semantic shift detection), usually in a monolingual setting using diachronic corpora (Tahmasebi et al., 2021; Tang, 2018; Kutuzov et al., 2018). Lately, the standard methods include contextual embeddings, but static embeddings remain a viable alternative.

In terms of the study of cross-lingual lexical semantic similarity, there have been a number of previous studies attempting to automatically extract pairs of true cognates and false friends from corpora or from dictionaries. Most methods are based either on orthographic and phonetic similarity, or require large parallel corpora or dictionaries (Inkpen et al., 2005; Nakov et al., 2009; Chen and Skiena, 2016; St Arnaud et al., 2017). There are few previous studies using word embeddings for the detection of false friends or cognate words, usually using simple methods on only one or two pairs of languages (Torres and Aluísio, 2011; Castro et al., 2018).

While the study of lexical semantic change detection has recently gained popularity in the NLP community, where automatically trained word sense representations are generally used to empirically measure and characterize semantic change, this has almost exclusively been done in monolingual settings. Uban et al. (2021) propose using cross-lingual semantic divergence between cognate words in a synchronic setting as a way to measure semantic shift from the original etymon to the present day words, and experiment using a small cognate database and static embeddings.

Starting from this idea, our study proposes a systematic analysis of false friends detection in Romance languages, including the following contributions: we compute a comprehensive analysis of cognates and borrowings semantic divergence on the most complete available dataset of cognate words and borrowings in Romance languages. We propose a benchmark for false friends detection correction for words in the Romance languages including (for the first time, to our knowledge) both deceptive cognates and borrowings, and publish the obtained lexicon of deceptive cognate and borrowing pairs (which we make freely available to use for research purposes) which can serve as exhaustive false friends lists for any cognate pair in the studied Romance languages.

Thus, in the current study we aim to address the following research questions:

RQ1. Can the synchronic study of current meanings of words with common etymology help us

measure their semantic change over time?

RQ2. How reliable are automatically generated multilingual word sense representations for detecting cross-lingual semantic change and for automatic false friends detection and correction?

2 Cognates and Borrowings Dataset and Corpora

Cognates and Borrowings Dataset We perform our analyses on related word pairs extracted from the most comprehensive database of related words in cognate languages up to date, sourced from etymological dictionaries and manually curated, RoBoCoP (Dinu et al., 2023). As a source of cognate word pairs, we use the freely available subset ProtoRom (Dinu et al., 2024a), a database of cognate tuples and etymons in the five Romance languages, with 19,222 entries (tuples with at least 2 cognates). We extract borrowings from the original RoBoCoP database, totaling 46,490 borrowing pairs across Romance languages pairs (Dinu et al., 2024b).

Word Embeddings Corpora For our computational experiments, we rely on word embeddings as models of meaning representation. In order to compare the effect of the corpus use to train the embeddings, we experiment with three different parallel aligned corpora to extract embeddings:

- Wikipedia[1]
- *Europarl*, a standard parallel corpus with aligned sentences including the Romance languages, based on proceedings of the European Parliament (Koehn, 2005),
- *RomCro2.0*, a recent parallel corpus including more general language sourced from literary works written in various original languages and translated in Romance languages and Croatian (Mikelenić et al., 2024),

False Friend Annotation For each language pair, we provide a set of labels derived from Open Multilingual WordNet (Bond and Foster, 2013), as well as manual annotations on a sample of the word pairs.

Multilingual WordNet (Bond and Foster, 2013) is organized as a multilingual semantic network with links between synsets across languages. Two cognates or borrowings are considered "true cognates/borrowings" if they occur in the same synset across languages and "deceptive cognates/borrowings" otherwise). While WordNet has the advantage of allowing automatic extraction of labels for any word in the vocabulary in theory, its coverage on our dataset is poor. For this reason, we manually annotate a subset of approximately 10% of word pairs for each language pair out of the vocabulary covered by the corpora used (but no more than 100 and no less than 10 examples per language pair, separately for cognates and borrowings), to serve as a more reliable ground truth as described further in this section. More statistics on the number of cognates and coverage in WordNet and corpora are reported in the Appendix (Section A.2).

Linguists specialized in each of the Romance languages under study manually annotated the pairs of cognates and borrowings as true or deceptive, a process that involved two steps: (1) consulting the main monolingual dictionaries for each language and bilingual dictionaries of each pair of languages and (2) checking for meaning and usage matches by activating machine translation or accessing parallel text sites such as Linguee. The clearest situation of deceptive cognates/borrowings is when the lexicographic definitions of the two corresponding words do not coincide (their semantic areas do not overlap): e.g. Es. *ración* 'part or portion of food that is given to humans or animals' vs Ro. *rațiune* 'reason' // It. *salire* 'climb up'/ Es. *salir* 'get out'. In such a case, we clearly have a pair of deceptive cognates, which will be marked with 1. The second situation is when the main meaning of a word coincides with one of the secondary meanings of its pair: e.g. Ro. *radia* 'to emit rays of light, heat, sound waves' / It. *raggiare*, for which the main acceptation is 'to design with ray-like figures a sheet, a fabric', but which includes in its semantic area the secondary meaning of 'emanate or reflect light rays'. A situation as such in which the semantic areas intersect but do not completely overlap has been marked with 2. The third situation encountered is when the first definition of the word in language A coincides with the first definition of its counterpart in language B; here it would seem, at first sight, that we have the ideal situation where the two words correspond perfectly. However, in cross-linguistic translation it emerges that the two terms are not equivalent, i.e., they would not be used in the same context: for example, It. *valigia* is the main word for 'luggage'; Es. *valija*, although it appears in lexicography with the same main mean-

[1] https://huggingface.co/datasets/wikimedia/wikipedia

ing, would never be used in everyday speech in contexts where an Italian uses *valigia*. In parallel contexts, a Spaniard would naturally use *maleta* or, in a more formal context, *equipaje*. This kind of situation has also been marked with a separate label and excluded from the computational experiments.

3 Methodology

In order to identify pairs of deceptive examples from corpora of cognate and borrowing pairs we design several approaches for classification and propose a method for correction. Our methods rely on language-aligned static word embeddings, contextual embeddings extracted with multilingual semantically-aligned transformer models, and combinations of both. The employed static embeddings are a pre-trained set gathered from Wikipedia, while for the contextual ones we experiment with contexts gathered from three corpora (Wikipedia, EuroParl, and RomCro). The correction method attempts to combine the knowledge of large portions of the languages' vocabularies provided by static embeddings and the contextual knowledge of a smaller subset of words provided by the transformers. In other words, we design a two-step pipeline in which static embeddings gather the closest N neighbors, while the contextual ones are used to re-rank them.

3.1 Word Meaning Representations

Our proposed algorithms rely on word embeddings to measure semantic distances based on embedding distances, using two different embedding algorithms:

- contextual embeddings extracted from a BERT transformer pretrained on a multilingual sentence similarity task for optimizing sentence representations, based on a Sentence-BERT architecture (Reimers and Gurevych, 2019) [2], as well as the multilingual transformer `xlm-roberta-base` (Conneau et al., 2019) for a subset of the experiments

- static FastText embeddings (Bojanowski et al., 2016), which have been previously used successfully for cognate semantic divergence measures (Uban et al., 2019)

[2]https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

Contextual embedding representations In order to extract unique vectorial representations for each cognate and borrowing from the pretrained sBERT model and the three corpora, we first identify each target word in our database in the corpus, based on their stems (obtained using the Snowball stemmer). We obtain for each cognate/borrowing a set of embeddings corresponding to each occurrence in the corpus (including potentially different senses of the word), and experiment with three different methods for computing distances between cognates based on the sets of their corresponding embeddings, inspired from the best solutions proposed in (Periti and Tahmasebi, 2024):

- mean distance: a simple dimension-wise average of the embeddings is computed to obtain unique representations per cognate, then cosine similarity is used to compute distances,

- JSD: embedding clusters for each cognate are generated using affinity propagation and cosine distance, the Janson-Shannon divergence is computed between the clusters as a distance metric between cognates,

- WID: embedding clusters for each cognate are generated independently, and cluster centers are computed using simple averaging, then the distance between clusters is computed as the cosine distance between cluster centers.

Static embedding representations In order to obtain comparable embedding representations across languages, it is necessary to align the monolingual spaces into a multilingual space. In the case of Wikipedia, we use pretrained prealigned static embeddings (Bojanowski et al., 2016; Joulin et al., 2018). For the smaller corpora for which there are no pretrained aligned embedding spaces across languages, we experiment with aligning the obtained spaces in order to obtain a single embedding vector space across languages (one for each corpus) based on a small bilingual dictionary curated manually, following the method in Lample et al. (2017). Preliminary results on these corpora show that false friend detection performance using static embeddings is very poor (due to their much smaller size compared to Wikipedia), so we exclude them from further experiments using static embeddings.

3.2 False Friends Detection

In the first phase of our experiments, we implement false friend detection as a binary classifier based

simply on cosine distance scores in the different embedding spaces. A distance threshold is optimized on the WordNet labels and used to obtain binary predicted labels for the manually curated test sample.

3.3 Reranking and False Friend Correction

The algorithm based on embedding distances inherently allows for generating "translations" of cognate or borrowing pairs that are identified as false friends based on closest neighbors in the embedding space. We propose that these can be interpreted as "false friend correction" suggestions by the algorithm in cases where the cognate or borrowing is not the correct translation of the original word.

In a second stage of our experiments, we extend our algorithm to produce a more flexible output, as a ranking of suggested translations for a given cognate or borrowing. This allows us to frame the problem of identifying false friends as a retrieval problem by evaluating where true cognates and borrowings (analogous to relevant documents in our case) are positioned in the rankings provided by the algorithm in comparison to false friends.

False friend correction was proposed in Uban and Dinu (2020) based on retrieving the closest neighbor in the static embedding space.

We extend this method by combining static and contextual embedding spaces. We use the embedding representations extracted based on Wikipedia (trained on Wikipedia in the case of static embeddings, and extracted based on Wikipedia contexts in the case of contextual embeddings), which perform best in the first stage of the experiments based on the F1-scores obtained for false friend detection. While the retrieval of nearest neighbors in static embedding space is reasonably easy to compute, since distances to every word in the vocabulary can be computed and compared, in contextual embedding space it computationally expensive to compute representations and distances for every word in the vocabulary. In order to combine the two types of embedding representations, we first generate a ranking of correction suggestions based on cosine distance in the static embedding space, which in a second step we refine by reranking based on contextual embeddings. In this way, we limit the computational effort by only computing contextual representations and distances for a subset of the vocabulary - the candidates produced by the static embedding model.

We propose two different methods of incorporating contextual embedding information in the ranking mechanism, as described further in this section.

We additionally extend the algorithm to output multiple suggestions, in order of relevance, and associated with a score derived from distance in embedding space. As evaluation metrics, we use:

- $Cov@k$ (coverage in top k or top k accuracy), is a relaxed metric which computes the percentage of input words for which the k-best output list contains the correct solution.

- Mean reciprocal rank (MRR) (Voorhees et al., 1999) is the multiplicative inverse of the rank of the first correct answer. Given an input word, the higher the position of its correct answer in the output list, the higher the MRR.

- Bpref (De Beer and Moens, 2006), a preference-based information retrieval measure that considers whether relevant documents (in our case, true cognates or borrowings) are ranked before irrelevant ones (false friends) .

Mask-based corrections reranking The first method relies on leveraging the masked language model task through a multilingual transformer, `xlm-roberta-base` (Conneau et al., 2019). Given a word in language A and a set of words in language B, that were retrieved as the closest words in the aligned static embedding spaces, we are going to approximate the probability that a word in B can replace the word in A in various contexts. For this, we pick the occurrences of the word in language A that were found in the Wikipedia dataset and run the MLM on those sentences with a mask covering this word. The returned logits represent the probability distribution of various tokens that can replace that mask. We average the logits obtained from all of the contexts and for a given word in B we calculate a score as the sum of probabilities for the sub-tokens that form the word in B. The set of words in B are then sorted based on these scores to provide a reranking.

Contextual distance-based corrections reranking The second method computes contextual embeddings using the aligned sBERT model for the occurrences of the word from language A and for the occurrences of all words in language B. For finding the contexts we again employ the Wikipedia

dataset. We then average these contextual embeddings for each word individually and compute the cosine similarity between the average embedding of the word in A and the average embeddings of the words in B. Based on these scores, the words in B are sorted and a reranking is provided.

4 Results

Results for the first stage of our experiment for false friend detection with the best method are shown in Figure 1. Table 1 displays the best F1 scores computed on the test set (the manually annotated examples) using the pre-trained static embeddings and the best contextual embeddings extracted from the Wikipedia dataset. We observe the highest level of semantic divergence in the case of borrowing pairs from Es in Ro. Also noticed during the manual annotation process, this divergence is due to the fact that Ro has borrowed words from Es with a specialized meaning, or with an applicability circumscribed to Hispanic realities, whereas in Spanish the words have a general meaning: e.g. Es. *alcázar* means 'fortress', whereas Ro. *alcazar* was borrowed with the highly specialized meaning of 'fortress built by Moors in Spanish cities in the Middle Ages': therefore, the two words will not be used in the same contexts.

We find that contextual embeddings lead to better deceptive cognate detection performance than static embeddings, best results obtained on Wikipedia and slightly better results on the smaller but more general RomCro2.0 corpus than on Europarl. Thus, more extensive pretraining and more sophisticated contextual representations might be more powerful for semantic representations of words than the traditional static embeddings. It is interesting to notice that the sBERT models trained to maximize sentence similarity can still induce some word-level alignment to produce comparable word embeddings in a multilingual space, without explicit alignment.

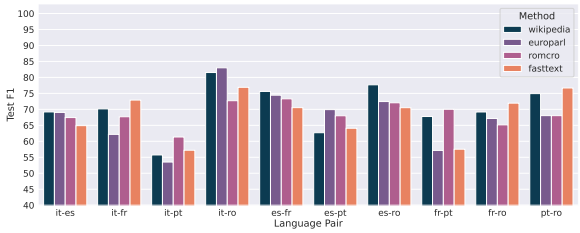
For the second stage of our experiment, results show that reranking is effective as a solution for improving on static embedding suggestions for deceptive cognate correction using the context based reranking method (based on contextual embedding nearest neighbors), whereas the mask based method degrades results in all cases. The effects of reranking are illustrated in Figure 2, showing an overall improvement in the Cov5 metric (where true cognates are the positive class) using the context based method. This suggests that the reranking method is effective at bringing up in the ranking more appropriate translations, which shows its effectiveness for suggesting potential false friend corrections. Figure 3 illustrates how the ranks of true and deceptive cognates comparatively change after reranking, showing that combining the two types of embeddings tends to have the desired effect of pushing true cognates towards the top of the ranking and pushing false friends towards the bottom. The same metrics computed for borrowings, including split by borrowing direction, can be found in the Appendix (Figures 4 and 5). Other metrics such as Bpref are not as affected by reranking, more details are reported in the Appendix (Section A.2, Figures 7 and 8).

5 Discussion

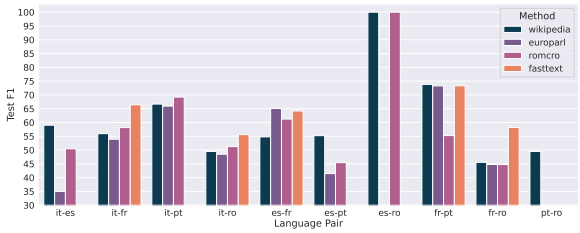
While some examples of deceptive cognates are observable to the naked eye, others are hard to detect even by linguists specialized in Romance languages. Among those which are obvious even to the non-specialized speaker, we mention the case of the Romanian word *aprinde* (meaning 'to ignite'), which creates deceptive cognate pairs with words from the Wester Romance languages, Fr. *apprendre*, It. *apprendere*, Es./Pt. *aprender* (all of them meaning 'to learn'). There are also several cases where the difference in nuance is small enough for even linguists to miss it, but it is identified correctly by automatic detection: e.g. linguists have marked the pair Sp. *marchar*/ Fr. *marcher* (where Es is borrowed from Fr) as true borrowings, given that the Spanish word is defined as "to go to some place", and the French one as 'to go from some place to another'. However, Fr. *marcher* actually means 'to walk', whereas in Spanish this concept is translated as 'caminar' or "andar". In other cases, polysemy can lead to ambiguous linguistic situations. One such case is that of the Romanian term *absolvi,* that can mean either 'to graduate' or 'to absolve'. Ro. *absolvi* is identified by the machine as true cognate in relation to the French term *absoudre* ('to absolve') and as false friend with the Portuguese term *absolver* ('to absolve').

Inconsistency can also be of morphological nature, as certain parts of speech do not coincide in certain contexts. This is the case of Pt. *insultar* (verb) – Ro. *insultă* (noun) and It. *regolare* (verb) – Ro. *regulă* (noun).

There are some cases where the machine did



(a) Deceptive cognates



(b) Deceptive borrowings

Figure 1: Deceptive cognate and borrowing detection results (threshold-based) using the best method (static and contextual embeddings based on Wikipedia).

	RO				IT				PT				FR				ES			
	St.	Cos	JSD	WID	St.	Cos	JSD	WID	St.	Cos	JSD	WID	St.	Cos	JSD	WID	St.	Cos	JSD	WID
RO	–	–	–	–	77%	77%	73%	**82%**	**77%**	74%	71%	75%	**72%**	68%	58%	69%	71%	69%	64%	**78%**
IT	**56%**	48%	49%	50%	–	–	–	–	**57%**	50%	47%	56%	**73%**	62%	57%	70%	65%	65%	61%	**69%**
PT	–	23%	**59%**	23%	–	**67%**	53%	62%	–	–	–	–	58%	65%	61%	**68%**	**64%**	62%	59%	63%
FR	**58%**	45%	44%	46%	**66%**	60%	41%	49%	73%	**76%**	60%	64%	–	–	–	–	71%	72%	63%	**76%**
ES	–	**100%**	44%	38%	–	55%	50%	**59%**	–	**55%**	54%	45%	**64%**	55%	55%	54%	–	–	–	–

Table 1: Deceptive cognate and borrowing detection performance using F1-score, threshold-based setting (cognates scores above the main diagonal and borrowing scores below the diagonal, averaged across the two directions for borrowing for each language pair), for each language pair and embedding distance metric, using the best static and contextual embeddings models (based on Wikipedia).

not mark a pair as false friends, although the corresponding words do not show total equivalence when it comes to language use: e.g. Pt. *pêlo* / Ro. *păr*; while Ro. refers to any kind of hair, Pt. is used mainly for body hair or animal fur, the term preferred for hair on the head being *cabelo.* In the line of language use, although Pt. *porco* ('pig') appears to be a true cognate of Fr. *porc*, in French the term refers mainly to pork meat, the preferred term for the animal being *cochon.*

The ranking-based evaluation of false friend correction suggestions proposed by the algorithm indicates that the generated rankings could constitute a viable solution for assisting a user or a linguist with finding the correct "translation" of a deceptive cognate or borrowing. Some examples of successful correction include Ro. *mică* / It. *mica*: both come from the Latin *mica* 'small piece', but Ro. was specialized as the adjective 'small', and It. has restricted its meaning to 'breadcrumb', therefore they are deceptive cognates. In this case, the machine proposes as equivalent for Ro. *mică* It. *piccola*. The pair Ro. *holeră*/ It. *collera* are deceptive cognates, both coming from Lat. *cholera* 'bile'; but, whereas the Romanian word has acquired the meaning of 'cholera', the Italian one, in this phonetic variant, means 'anger'. The algorithm replaces it with the true cognate of Ro. *holeră*, namely *colera* - the variant derived from the same etymon, but encapsulating the identical semantic evolution of the Romanian word.

There are also situations in which the proposed alternative is not semantically equivalent to the target word, but may be part of the same conceptual field, as used in the same phrases: for example, Ro. *rotund* 'round' and Es. *redondo* 'id.' are true cognates, although the contexts in which the Spanish word can be used are more diversified (the adverbial locution *en redondo* "categorically" became rather frequent in the last few years), which is why the computer finds a rather large distance between the two lexemes: as a consequence, it proposes as equivalent for Ro. *rotund* the Spanish word *aplanado* 'flat', no doubt from the increasing usage of the phrase "the world is flat", instead of the expected "round". (After improving the algorithm with re-ranking based on contextual embeddings, it more accurately proposes "redonda" as an alternative correction.)

6 Conclusions and Future Directions

We perform an exhaustive analysis of the semantic shifts across cognates and borrowings in Romance languages, based on a comprehensive and reliable dataset of Romance related words, using only synchronic data. While the synchronic approach is inherently limited in its ability to track historical trajectories, our results show that embedding-based semantic distances between cognates and borrowings based on contemporary language corpora can reflect semantic divergence between related words

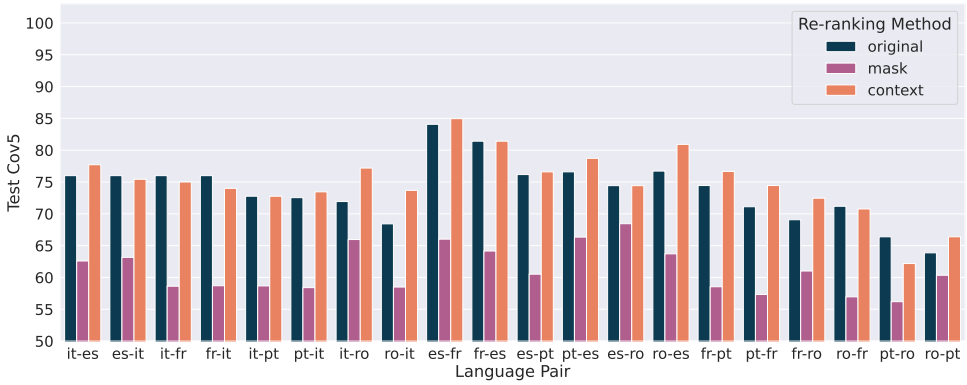


Figure 2: Effects of reranking correction for cognates suggestions based on Cov5.

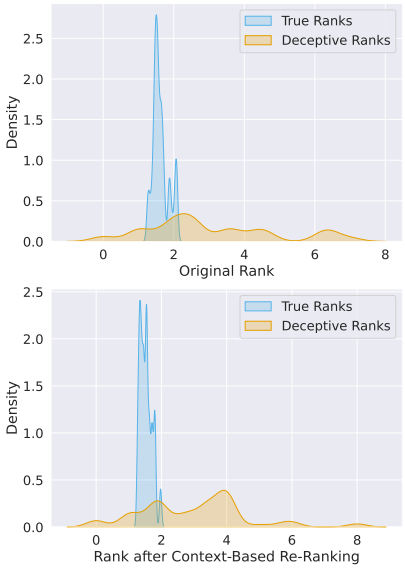


Figure 3: Distribution of false friend ranks before and after reranking.

in Romance languages, and successfully detect semantic false friends (deceptive cognates and borrowings). We propose metrics that allow us to define deceptiveness as a spectrum and view how similarly or how differently the meanings of modern day words have shifted (RQ1).

We compare different word representations and algorithms to obtain a benchmark for false friend detection and correction, and find that contextual embeddings based on transformers pretrained to optimize sentence similarity are the most useful for detecting deceptive cognates across Romance languages, and that combining static and contextual representations can be helpful for obtaining useful false friend "corrections" in a computationally effective manner (RQ2).

We publish the resulted distance scores for every cognate and borrowing pair in all languages as a freely available lexicon, which can be easily used to automatically extract all false friends for any Romance language pair, based on a customizable distance threshold depending on the application (even for words that are outside the vocabulary of electronic dictionaries such as WordNet).

At the technical level, in future work, it would be worthwhile to explore whether postprocessing the contextual embedding space to align word representations can lead to a more consistent cognate vector space and produce better results.

Furthermore, we aim to open the following research directions.

- While our method shows it is possible to detect semantic shifts based on synchronic corpora, in cases where the semantic distance between cognates/borrowings is significant, it could be a useful improvement to include diachronic corpora and track this divergence diachronically: starting from the source word, we can trace to what extent was the original meaning preserved, marking on a scale the level of divergence;

- For an accurate perspective, it is necessary to extend the range of meanings that we take into account in both the source and the target language, given that most words are polysemic: to use an example cited above, if we consider the relation La. *vindicare* 'to revenge' > Ro. *vindeca* 'to cure', we will not limit our view to the first meaning given by Latin dictionaries, but we will contemplate the whole range of meanings; we will thus notice that Ro. *vindeca* actually preserves one of the multiple Latin meanings, and does not create a new meaning. The proposed methods based on contextual embeddings already support poly-

semy, so a more nuanced model of semantic divergence could be implemented based on our method just by extending the annotation schema from binary to a scale of change;

- Going further into studying the properties of the discovered semantic shifts across languages, as well as the specificities of the kinds of words that tend to undergo semantic change, towards a systematical analysis of the laws of semantic divergence cross-lingually (Hamilton et al., 2016; Uban et al., 2021) is another important direction for future research;

- Beyond looking at semantic change as a one-dimensional spectrum, following this approach, we suggest that it would be useful to a propose a categorization of inherited or borrowed words according to the type of lexico-semantic process they have undergone: e.g. loss of the main meaning, change of the hierarchy of meanings within their semantic area, addition of a new meaning, etc. Beyond binary false friend detection, this framework would allow for a systematic analysis of subtypes of semantic change in words with common etymology.

Limitations

Some refinements of the data in the ProtoRom database, such as excluding words with multiple etymologies, which are not currently handled in the available version of the database, could be useful for a more accurate model of the linguistic phenomenon. Graphic issues can also lead to errors in detecting cognates. The Italian term *rio* ('river') and the Romanian term *rău* ('bad') were identified as false friends. However, the two terms don't share a common etymon – the first one comes from the Latin word *rivus* and the second one from lat. *reus* – and as such they can be considered chance false friends. A possible explanation for this pair resides in the graphical resemblance of *rău* ('bad') and *râu* ('river') and the misidentification of the diacritics on the letter *a, ă* and *â.*

Ethical Statement

There are no ethical issues that could result from the publication of our work. Our experiments comply with all license agreements of the data sources used. We make the contents of our package available for research purposes upon request.

Acknowledgments

All authors contributed equally to this research. This work was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization, and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and by Ministry of Research, Innovation and Digitization, CNCS- UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023-1701, within PNCDI IV.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Santiago Castro, Jairo Bonanata, and Aiala Rosá. 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 (VarDial 2018)*, pages 29–36.

Yanqing Chen and Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jan De Beer and Marie-Francine Moens. 2006. Rpref: A generalization of bpref towards graded relevance judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 637–638.

Liviu P Dinu, Ana Uban, Alina Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7610–7629.

Liviu P Dinu, Ana Uban, Alina Cristea, Ioan-Bogdan Iordache, Teodor-George Marchitan, Simona Georgescu, and Laurentiu Zoicas. 2024a. Verba volant, scripta volant? don't worry! there are computational solutions for protoword reconstruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6314–6326.

Liviu P Dinu, Ana Uban, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2024b. It takes two to borrow: a donor and a recipient. who's who? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6023–6035.

Pedro J Chamizo Dominguez and Brigitte Nerlich. 2002. False friends: their origin and semantics in some selected languages. *Journal of pragmatics*, 34(12):1833–1849.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL 2016*, pages 1489–1501.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of RANLP 2005*, volume 9, pages 251–257.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Wilhelm Meyer-Lübke. 1911. *Romanisches etymologisches wörterbuch*, volume 3. C. Winter.

Bojana Mikelenić, Antoni Oliver, and Marko Tadić. 2024. Expansion of the romcro corpus with texts in catalan. In *CLARIN Annual Conference Proceedings 2024*, pages 135–139. Barcelona: CLARIN.

Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of RANLP 2009*, pages 292–298.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. *CoRR*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying Cognate Sets Across Dictionaries of Related Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Lianet Sepúlveda Torres and Sandra Maria Aluísio. 2011. Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Ana Sabina Uban, Alina Ciobanu, and Liviu P Dinu. 2021. Cross-lingual Laws of Semantic Change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*. Berlin: Language Science Press.

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2019. Studying Laws of Semantic Divergence across Languages using Cognate Sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.

Ana Sabina Uban and Liviu P Dinu. 2020. Automatically Building a Multilingual Lexicon of False Friends With No Supervision. In *Proceedings of LREC 2020*, pages 3001–3007.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

A Appendix

A.1 Infrastructure and libraries

The experiments were performed on an RTX 2080 Ti GPU and a Ryzen 5 3600X CPU for a total of 72 hours.

Libraries used for embedding extraction, cognate and corpora preprocessing (extracting stems), synonym extraction based on WordNet, and distance metrics computation:

- keras==3.8.0
- keras-hub==0.18.1
- keras-nlp==0.18.1
- nltk==3.9.1

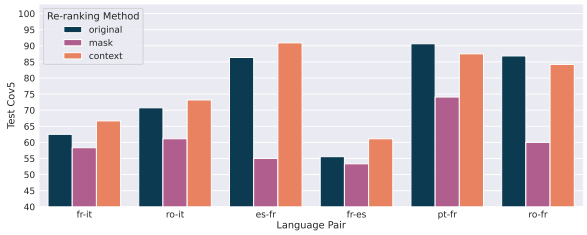


Figure 4: Effects of reranking of Cov5 for borrowings.

- scikit-learn==1.6.1
- scipy==1.13.1
- sentence-transformers==3.4.1
- spacy==3.7.5
- tensorflow==2.18.0
- tensorflow-datasets==4.9.7
- transformers==4.48.3
- and fasttext vector support based on https://github.com/babylonhealth/fastText_multilingual/.

Transformer models used:

- `distiluse-base-multilingual-cased-v2`: 135M parameters
- `xlm-roberta-base`: 279M parameters

Hyperparameters:

- maximum number of sampled occurrences for a word when computing contextual embeddings: 200
- occurrence matching was checked based on stem matching with and without unicode normalization (removing of accents)
- Affinity Propagation clustering was trained with the default hyperparameters provided by the scikit-learn library.

A.2 Additional Results

The number of cognates pairs for each language pair and coverage in WordNet for the cognates in our database is as follows:

- IT-ES: total: 3666, not in WordNet (WN): 1923
- IT-FR: total: 2172, not in WN: 918
- IT-PT: total: 10421, not in WN: 6479
- IT-RO: total: 2445, not in WN: 1143
- ES-FR: total: 4091, not in WN: 2196
- ES-PT: total: 4018, not in WN: 2131
- ES-RO: total: 5844, not in WN: 3340
- FR-PT: total: 2232, not in WN: 975
- FR-RO: total: 3416, not in WN: 1626
- PT-RO: total: 2545, not in WN: 1280

The coverage of cognates in the corpora used is as follows:

- RO: Total ProtoRom Words: 5522, Found in EuroParl: 3357 (60.79%), Found in Wikipedia: 5248 (95.04%)
- IT: Total ProtoRom Words: 7587, Found in EuroParl: 5576 (73.49%), Found in Wikipedia: 7431 (97.94%)
- ES: Total ProtoRom Words: 6361, Found in EuroParl: 5468 (85.96%), Found in Wikipedia: 6342 (99.70%)
- FR: Total ProtoRom Words: 3991, Found in EuroParl: 3160 (79.18%), Found in Wikipedia: 3952 (99.02%)
- PT: Total ProtoRom Words: 9107, Found in EuroParl: 5851 (64.25%), Found in Wikipedia: 8391 (92.14%)

	RO	IT	PT	FR	ES
RO	–	83%	68%	67%	72%
IT	48%	–	53%	62%	69%
PT	–	66%	–	57%	70%
FR	44%	54%	73%	–	74%
ES	-	35%	41%	65%	–

Table 2: False friends detection performance measured in F1-scores on the test set (cognates above the main diagonal and borrowings below the diagonal) for each language pair and the best embedding distance metric, using contextual embeddings on the Europarl corpus.

	RO	IT	PT	FR	ES
RO	–	72%	68%	65%	7%2
IT	51%	–	61%	67%	67%
PT	–	69%	–	70%	68%
FR	44%	58%	55%	–	73%
ES	100%	50%	45%	61%	–

Table 3: False friends detection performance measured in F1-scores on the test set (cognates above the main diagonal and borrowings below the diagonal) for each language pair and the best embedding distance metric, using contextual embeddings on the Romcro corpus.

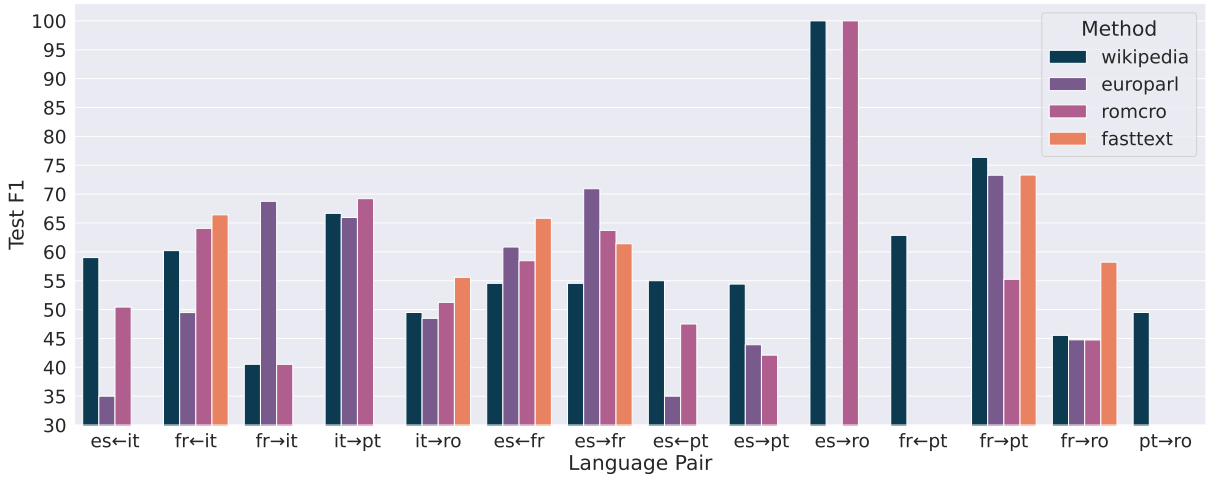


Figure 5: Effects of reranking of F1-score based on first prediction for borrowings, separately for the two borrowing directions.

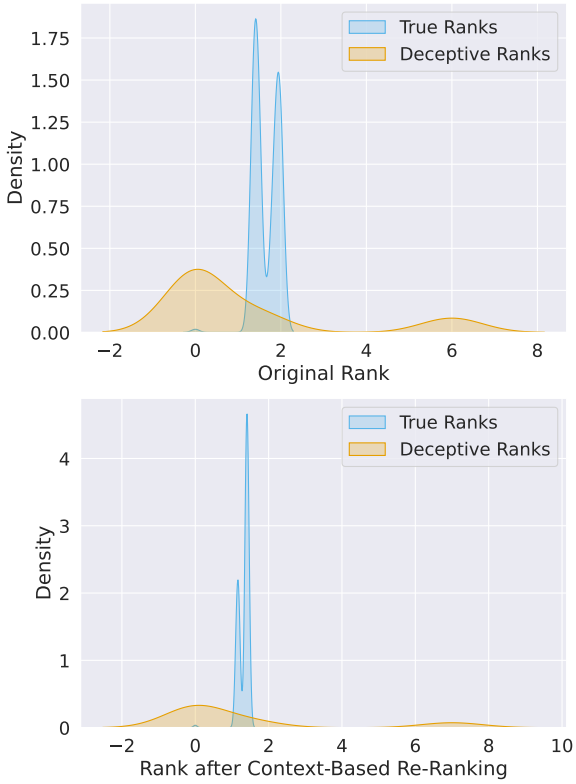


Figure 6: Effects of reranking on MRR for borrowings.

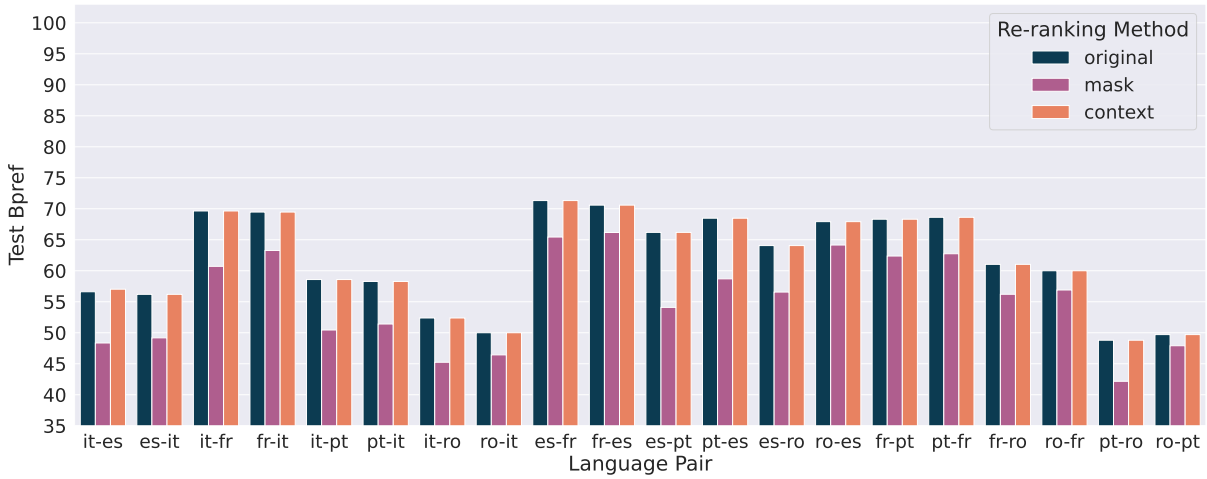


Figure 7: Effects of reranking on Bpref for cognates.

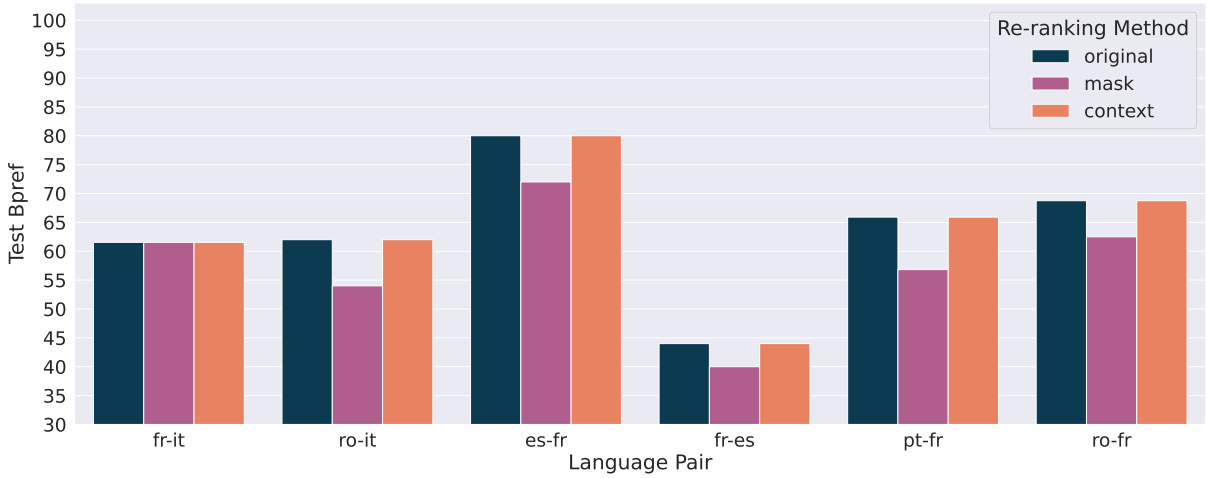


Figure 8: Effects of reranking on Bpref for borrowings.

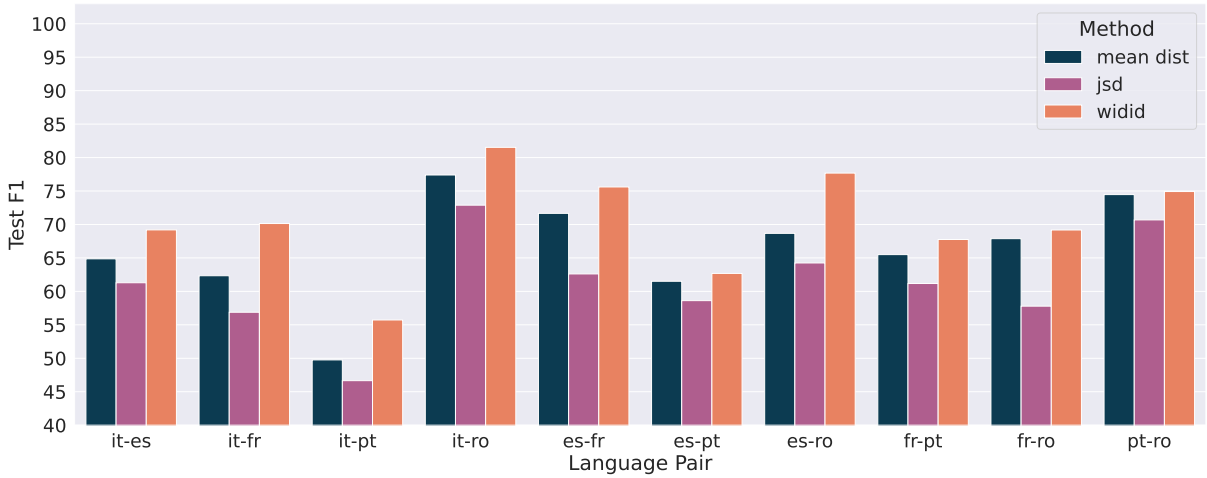


Figure 9: Comparison of different distance metrics employed for deceptive cognates detection using contextual embeddings extracted from Wikipedia. We report F1 scores on the test set.

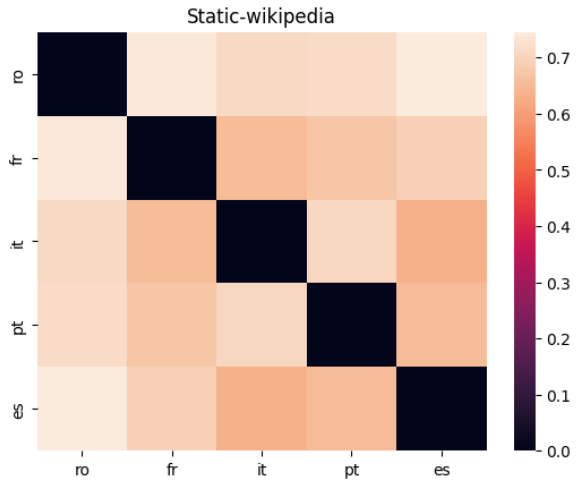


Figure 10: Semantic divergence between Romance languages based on average cluster distance on contextual embeddings on the Wikipedia corpus

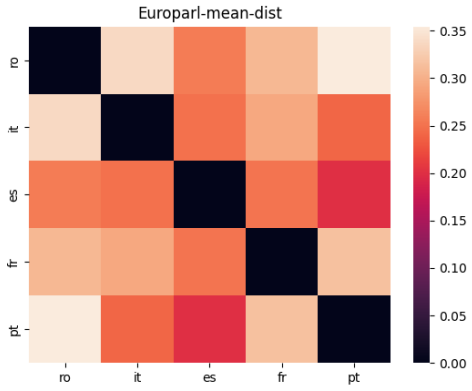


Figure 11: Semantic divergence between Romance languages based on static embeddings trained on Europarl